Can Humans Add Value When Algorithms Outperform Them?*

Nikhil Agarwal Simon Hodson Ashesh Rambachan
Yucheng Shang†

November 3, 2026

Abstract

Algorithms often outperform human decision makers in high-stakes settings. We study what can be robustly learned about human and algorithmic decision making from these comparisons and the data that underlie them. We develop a nonparametric framework in which each choice applies a threshold to the decision maker's prediction, and we clarify the implicit restrictions on preferences under which the comparisons establish that humans have worse predictive skill. We further show that the data behind these comparisons are informative about how to combine humans and algorithms: they can test whether the decision maker has private information, and provide bounds on the value of eliciting the decision maker's input and the value of AI assistance. In New York pretrial release, the algorithm outperforms 95 percent of judges at their release rates, yet about a quarter of judges hold private information, including judges the algorithm outperforms, and giving judges the algorithm's prediction lowers their release-misconduct rates by at least 5.9 percentage points on average.

*All errors are our own.

†Agarwal: Department of Economics, MIT and NBER (agarwal3@mit.edu); Hodson: Blueprint Labs, MIT (simon600@mit.edu); Rambachan: Department of Economics, MIT (asheshr@mit.edu); Shang: Department of Economics, MIT (ycshang@mit.edu).

1 Introduction

Algorithms that predict outcomes from data now inform high-stakes decisions, from pretrial release to medical diagnosis to hiring (Kleinberg et al., 2017; Mullainathan and Obermeyer, 2021; Li, Raymond and Bergman, 2020). A common finding is that these algorithms outperform the humans who make these decisions (Kleinberg et al., 2017; Mullainathan and Obermeyer, 2021; Angelova, Dobbie and Yang, 2023; Liu et al., 2019). Holding the rate of decisions fixed, the algorithm classifies better: it would release the same share of defendants with less pretrial misconduct, or test the same number of patients and detect more of those at risk. These comparisons of predictive performance increasingly guide how decisions are allocated between humans and algorithms.

In this paper, we study what can be robustly learned about human and algorithmic decision making from these comparisons and the data that underlie them. If a decision maker's performance lies below the algorithm's frontier, are they worse at predicting the outcome? Even if they underperform the algorithm, could they still add value to its prediction? Answering either question requires measuring the value of the decision maker's information separately from their preferences. Yet we observe only their choices, which combine the two. A judge who releases a defendant the algorithm flags as risky may assess the risk differently, or may weigh the costs of detention against the costs of misconduct differently. A decision maker's choices may therefore fall short of the algorithm not because they predict worse, but because they weigh the two errors differently. The core challenge is to somehow separate what the decision maker knows about the outcome from how they weigh potential errors using only their observed choices.

We develop an econometric framework that uses the data behind these human-algorithm comparisons to bound the predictive value of the decision maker's information. The bounds let us revisit the questions that matter for the design of systems that combine humans and algorithms: whether the decision maker's predictions compare favorably with the algorithm's, whether they hold private information the algorithm lacks, and what is gained by combining their information with the algorithm's prediction. We apply the framework to pretrial release in New York State, where we pair the quasi-random assignment of judges to cases with the framework to bound each judge's predictive skill and the value of combining their information with the algorithm's prediction. We also have access to restricted court records for New York State covering arraignments from 2020 to 2026, which record each case in more detail. A future draft will apply our analysis to that sample.

Our framework analyzes a nonparametric, behavioral model of how a decision maker turns their information into choices. Each choice applies a threshold to the decision maker's

prediction of the outcome. The prediction summarizes everything in their choices that is correlated with the outcome, and the threshold collects everything else, which we refer to loosely as their preferences. The model captures many intuitions about human decision making in high-stakes settings. The decision maker may draw on features the algorithm never sees, such as a judge's interaction with a defendant at the hearing, in forming their predictions. Their prediction need not be accurate, so the decision maker may make systematic prediction mistakes (Rambachan, 2024). Their threshold may vary across individuals, so the same prediction can lead to different choices through preferences or noise (Kahneman, Sibony and Sunstein, 2021). On its own, the model has no testable implications: any choices are consistent with some prediction and some threshold. It nevertheless lets us pose each of our questions precisely and ask what the data reveal about its answer.

We then ask how much sharper the answers become under additional restrictions on the decision maker's preferences. Monotone choice behavior requires that the decision maker select an individual less often the higher their prediction of the outcome, which holds whenever their threshold varies for reasons unrelated to their prediction. A fixed threshold requires a single cutoff for every individual, which is imposed in existing structural models of expert decision making (Chan, Gentzkow and Yu, 2022; Arnold, Dobbie and Hull, 2022; Ribers and Ullrich, 2023; Rambachan, 2024). We show that both preference restrictions are testable through one comparison, whether the decision maker selects better than random: the individuals they select must be less risky on average than those they do not.

Our first set of results shows that the common comparison measures predictive skill only under a fixed threshold, and that relaxing this restriction changes the verdict for many judges. The literature compares a decision maker's performance along two dimensions against the algorithm's frontier, true positive and false positive rates or, in pretrial release, release rates and release-misconduct rates. We show that this comparison is valid only under a fixed threshold: when the threshold varies across individuals, the decision maker's choices lie strictly inside the frontier of their own prediction, and the gap to the algorithm mixes predictive skill with preferences. Under the model alone, the data can never show that the algorithm outperforms the decision maker's prediction, because any choices are consistent with a prediction that nearly reveals the outcome. Under monotone choice behavior, the data instead bound the decision maker's information frontier, from the frontier of their own choices to an upper bound that lies strictly below perfect prediction whenever, across the algorithm's predictions, the decision maker selects some individuals who go on to have the outcome more often than others who do not, so the data can still show that the algorithm outperforms the decision maker.

In New York, the common comparison finds that the algorithm beats 95 percent of

judges at their own release rates. Under the model alone, the data cannot rank 99.4 percent of the judges with cases in every decile of predicted risk against the algorithm. Only 61.6 percent of these judges select better than random. The remaining 38.4 percent release riskier defendants than they detain, so neither preference restriction can rationalize their choices, and the data alone cannot say why. Among all of them, the algorithm beats 33.7 percent under monotone choice behavior and 61.0 percent under a fixed threshold, so relaxing the fixed threshold cuts the share by about half, while only 0.6 percent beat the algorithm under every restriction. Monotone choice behavior therefore remains informative where a fixed threshold is implausible.

Our second set of results shows that the data underlying comparisons of humans and algorithms can be used to robustly test whether the decision maker has private information. The decision maker has private information if their prediction is informative about the outcome among individuals with the same prediction of the algorithm. They have private information at a prediction of the algorithm if and only if their choice is correlated with the outcome among the individuals with that prediction. This requires no restriction on the decision maker's preferences, and evidence of private information is robust: a correlation between the choice and the outcome cannot be rationalized without an informative prediction.

Many judges have private information, including judges the algorithm outperforms. We reject the null hypothesis of no private information for 35 percent of judges before adjusting for multiple testing and for 25 percent after adjusting for the false discovery rate, and the share is similar among the judges the algorithm beats, among those the bounds cannot rank, and among those who select worse than random. Private information is most prevalent in the extreme deciles of predicted risk, where the algorithm is most confident. This contrasts with work finding that AI assistance helps most when the algorithm's predictions are extreme, but that decisions are better automated once the algorithm is confident enough (Agarwal et al., 2023; Agarwal, Moehring and Wolitzky, 2025).

Our third set of results bounds the value of eliciting the decision maker's input to a designer who holds the algorithm's prediction and can elicit either the decision maker's choice or their prediction. The value of eliciting the choice is identified, and it exceeds the value of the algorithm alone, for some weight on the two classification errors, exactly when the choice is correlated with the outcome at some prediction of the algorithm. We further show that this identified value is a lower bound on the value of eliciting the prediction itself. The prediction summarizes everything the decision maker knows about the outcome, and the data record only the choice it produces, yet the choice alone guarantees a floor on what the designer gains from the prediction. Under a fixed threshold, the upper bound is

generally informative; as a result, under the restriction that justifies comparing humans and algorithms, the same data say much more about how to combine them.

In New York, at the judges' release rates, eliciting their choices lowers the release-misconduct rate by at least 0.9 percentage points from the algorithm's 26.0 percent. We cannot rule out that eliciting their prediction would lower it by 5.7 percentage points, more than a fifth of the algorithm's rate, under every restriction. Keeping judges in the loop can therefore be valuable even when the algorithm outperforms them.

Finally, we bound the value of AI assistance, the gain to the decision maker from observing the algorithm's prediction when the decision stays with them, at the weights their own threshold implies. Under the model alone and under monotone choice behavior, the data guarantee no gain, because the decision maker's threshold may already sit where the algorithm's prediction would change nothing. Under a fixed threshold, the data guarantee a positive gain whenever the individuals the decision maker selects at some prediction of the algorithm are riskier than those they do not select at another. In New York, under a fixed threshold, the algorithm's prediction lowers the judges' release-misconduct rates at their release rates by at least 5.9 percentage points on average and by at most 10.7, a fifth and a third of their rates. Once again, under the assumption justifying the common comparison, the same data can be used to bound the value of AI assistance to the decision maker.

Related Literature: A large literature compares human decisions with algorithmic predictions using realized choices and outcomes, and usually finds that algorithms do better on aggregate metrics of performance (Kleinberg et al., 2015, 2017; Mullainathan and Obermeyer, 2021; Ribers and Ullrich, 2023; Hoffman, Kahn and Li, 2017), a finding we replicate in our setting. Chan, Gentzkow and Yu (2022) compare decision makers with one another using similar metrics, so many of our results apply to those comparisons as well. Rambachan (2024) asks a related but different question, whether decision makers make systematic prediction mistakes, under a fixed threshold. Alur et al. (2023) test whether experts' decisions carry information beyond the available features, which is close to our test for private information. We analyze what the exercises in this literature imply about the comparison of predictive skill and their implications for the value of human input.

A second literature uses the quasi-random assignment of judges to study pretrial release decisions, in particular whether judges are biased against Black defendants (Arnold, Dobbie and Yang, 2018; Arnold, Dobbie and Hull, 2022), and related work measures bias in other settings (Dobbie et al., 2021; Baron et al., 2024). We use the same judge instrument to address the selective labels problem, but instead consider the quality of judges' predictions and their thresholds for choice.

Our results also bear on work on algorithmic triage and the design of AI assistance. Work

on triage and learning to defer decides which cases to route to the human (Raghu et al., 2019; Mozannar and Sontag, 2020), and related work studies how to combine human judgment with algorithmic predictions (McLaughlin and Spiess, 2023; Alur, Raghavan and Shah, 2024; Agarwal et al., 2023; Agarwal, Moehring and Wolitzky, 2025). Empirical work also studies how decision makers use algorithmic advice in practice (Angelova, Dobbie and Yang, 2023; Stevenson and Doleac, 2024; Imai et al., 2023). Many of these papers compare performance like the papers cited earlier and often allocate decisions using measured performance. This approach misses part of the value of human input. Eliciting decision makers' predictions rather than their choices (as in Agarwal et al., 2023; Agarwal, Moehring and Wolitzky, 2025) would recover their information frontier, which these allocation decisions need.

2 A Behavioral Model of Human Decisions

A decision maker uses their information about an unknown outcome to make choices for many individuals. A predictive algorithm summarizes each individual's recorded characteristics into a prediction of the same outcome. We study three questions. First, what do the decision maker's choices reveal about how well their information predicts the outcome compared with the algorithm? Second, would a designer who observes the algorithm's prediction gain by also eliciting the decision maker's choice or their information? Third, would the decision maker gain by observing the algorithm's prediction?

2.1 Setting and Observable Data

The decision maker selects a binary choice $c \in \{0,1\}$ for each individual. Each individual is summarized by characteristics $x \in \mathcal{X}$ and a binary outcome $y^* \in \{0,1\}$ that is unknown at the time of the decision. A predictive algorithm $m\colon \mathcal{X} \to [0,1]$ maps the characteristics into a prediction $m(x) \in [0,1]$ of the outcome. We treat the algorithm as known and fixed, and the decision maker need not observe its predictions. We assume the characteristics have finite support, so the algorithm's predictions also have finite support $\mathcal{M} \subset [0,1]$. We observe the algorithm's prediction M, the decision maker's choice C, and the outcome Y^*. The observable data are the joint distribution $(C, M, Y^*) \sim P(\cdot)$, which we assume is known to focus on the identification challenges in this setting. Our framework can be understood as conditional on coarse groups of individuals, in which case it concerns the decision maker's choices across cases within a group.

To make this concrete, we illustrate how a large class of empirical applications, known as *screening decisions*, maps into this setting.

Example: pretrial release A judge decides whether to release ($C = 1$) or detain ($C = 0$) each defendant awaiting trial (e.g., Kleinberg et al., 2017; Arnold, Dobbie and Hull, 2022;

Rambachan, 2024). The outcome Y^* is whether the defendant would commit pretrial misconduct (fail to appear in court or be rearrested if released). The characteristics record the defendant's demographics, current charges, and prior arrests and convictions. The algorithm maps these characteristics into a prediction M, as in the machine learning predictions of Kleinberg et al. (2017); Rambachan (2024) or the pretrial risk assessments studied by Angelova, Dobbie and Yang (2023). ▲

Example: medical testing and diagnosis A doctor decides whether to conduct ($C = 1$) or forgo ($C = 0$) a diagnostic test for each patient (e.g., Abaluck et al., 2016; Chan, Gentzkow and Yu, 2022). After an emergency room visit, for instance, a doctor decides whether to order a stress test for patients who may have had a heart attack (Mullainathan and Obermeyer, 2021). The outcome Y^* is whether the patient had a heart attack. The characteristics record the patient's demographics, reported symptoms, and prior medical history. The algorithm maps the patient's electronic health record into a predicted probability M of a heart attack. ▲

Example: hiring A hiring manager screens job applicants, deciding whether to interview ($C = 1$) or pass over ($C = 0$) each applicant (e.g., Hoffman, Kahn and Li, 2017; Cowgill, 2018; Li, Raymond and Bergman, 2020). The outcome Y^* is whether the applicant would be hired if interviewed. The characteristics record the applicant's demographics, education, and work history from the resume. The algorithm maps these characteristics into a predicted probability M that the applicant is hired. ▲

Such screening decisions are a leading class of "prediction policy problems" (Kleinberg et al., 2015; Mullainathan, 2025). Other examples include loan approvals (e.g., Fuster et al., 2021; Dobbie et al., 2021) and child welfare screenings (e.g., Baron, Ghosh and Lombardo, 2026; Baron et al., 2026).

As notation, we write $p(m) := P(M = m)$ and $p_{cy}(m) := P(C = c, M = m, Y^* = y)$. We write the outcome rate as $\mu := P(Y^* = 1)$ overall, $\mu(m) := P(Y^* = 1 \mid M = m)$ at prediction m, and $\mu_c(m) := P(Y^* = 1 \mid C = c, M = m)$ at prediction m given the choice $c \in \{0, 1\}$. We assume that the algorithm's prediction takes at least two values and that $p_{cy}(m) > 0$ for all $m \in \mathcal{M}$ and $(c, y) \in \{0, 1\}^2$.

2.2 The Behavioral Model

We model the decision maker's choices as arising from a threshold rule, which compares their prediction about the outcome against a threshold that summarizes their preferences. The behavioral model is a joint distribution over the algorithm's prediction, the outcome, their prediction, and their threshold that satisfies three conditions.

Definition 1. A *threshold model* is a joint distribution $(M, Y^*, H, T, C) \sim Q(\cdot)$ with $H \in [0,1]$ and $T \in (0,1]$ satisfying

i. Threshold Rule: $C = \mathbf{1}\{H \leq T\}$ almost surely.

ii. Independence: $T \perp\!\!\!\perp (M, Y^*) \mid H$.

iii. Calibration: $Q(Y^* = 1 \mid H) = H$.

The decision maker's choices are *consistent with* a threshold model $Q(\cdot)$ if $Q(M = m, C = c, Y^* = y) = p_{cy}(m)$ for all $m \in \mathcal{M}$ and $c, y \in \{0,1\}$.

Under Definition 1, the threshold T is independent of the algorithm's prediction and the outcome given H, so H summarizes all variation in the decision maker's choices that is correlated with the outcome. We therefore call H the decision maker's "prediction." Their prediction may draw on the characteristics or on features of each individual that are never recorded; for example, judges interact with defendants during the pretrial release hearing, but these interactions are not recorded. Calibration only requires that the outcome occurs at rate h among the individuals for whom they predict h; this places their prediction on the scale of the outcome but does not require it to be accurate: it allows for systematic prediction mistakes (Rambachan, 2024), and a prediction equal to the outcome rate μ for every individual is calibrated. The threshold T collects the remaining variation in their choices that is unrelated to the outcome given their prediction. We refer to T loosely as their "preferences," although it may also include noise in how the same prediction becomes different choices (Kahneman, Sibony and Sunstein, 2021). Altogether, Definition 1 attributes every factor driving their choices that is correlated with the outcome to their prediction, and everything else to their threshold.[1]

Since the algorithm's prediction and the outcome are observed, consistency requires $Q(M = m, Y^* = y) = P(M = m, Y^* = y)$ for all $m \in \mathcal{M}$ and $y \in \{0,1\}$; everything else Definition 1 imposes concerns their prediction and their threshold.

Lemma 1. *For every joint distribution $P(\cdot)$ of (C, M, Y^*), there is a threshold model that is consistent with the decision maker's choices.*

On its own, the threshold model has no testable implications. For example, the following threshold model rationalizes any choices: the decision maker's prediction equals the outcome

[1]This attribution is a normalization. In a structural model in which the decision maker's subjective probability of the outcome and their preferences are formed from what they observe and from shocks that carry no information about the outcome, the data do not separate the two on their own: either can absorb all of the variation in their choices that is correlated with the outcome. See Appendix B.1 for further discussion.

among the individuals they select and equals the outcome rate among those they do not select; because the threshold model leaves unrestricted how their threshold varies with their prediction, a threshold can then be chosen that reproduces their choices.[2]

Many threshold models are consistent with the decision maker's choices. The *identified set* $\mathcal{Q}(P)$ is the collection of threshold models $Q(\cdot)$ that are consistent with the decision maker's choices. We next introduce additional restrictions on their threshold that may shrink the identified set and produce testable implications.

Assumption 1 (Monotone Choice Behavior)**.** $Q(C = 1 \mid H = h)$ is nonincreasing in h.[3]

Assumption 2 (Fixed Threshold)**.** $T = t$ almost surely for some $t \in (0, 1]$.

Under the threshold model, $Q(C = 1 \mid H = h) = Q(T \geq h \mid H = h)$ is the probability that their threshold is at least h among the individuals for whom they predict h. Monotone choice behavior (Assumption 1) restricts how their threshold varies with their prediction: the threshold may rise with the prediction, but not so quickly that individuals with higher predictions are selected more often. The restriction is mild: it holds, for example, whenever their threshold varies for reasons unrelated to their prediction, such as noise or the judge's docket that day.[4] In the pretrial example, a judge exhibits monotone choice behavior if they release defendants less often the higher their predictions of a defendant's risk, however much noise or heterogeneity there is in how they do so. Under a fixed threshold (Assumption 2), a single cutoff converts their prediction into a choice for every individual, and so all variation in their choices must be driven by variation in their predictions. Existing structural models of expert decision making impose a fixed threshold together with calibrated beliefs (Chan, Gentzkow and Yu, 2022; Arnold, Dobbie and Hull, 2022; Ribers and Ullrich, 2023; Rambachan, 2024).

Remark 2.1. The threshold model can be interpreted as expected utility maximization (Appendix B.1). Suppose the decision maker holds a subjective belief that the outcome occurs and incurs a loss from each of the two errors they can make, selecting an individual for whom the outcome occurs or passing over one for whom it does not. Maximizing expected utility is then a threshold rule in their belief, with the threshold equal to the ratio of the

[2]Without calibration the threshold model has weaker conclusions in general: in addition to the flexibility in their threshold, the values of their prediction H may then be chosen freely.

[3]Formally, $Q(C = 1 \mid H = h)$ must be almost surely nonincreasing and left-continuous, and it must equal to one at $h = 0$. Appendix B.2 states the restriction formally.

[4]A stronger restriction is that the threshold is independent of the algorithm's prediction, the outcome, and their prediction, $T \perp\!\!\!\perp (M, Y^*, H)$. This implies monotone choice behavior and includes a fixed threshold as a special case. In Appendix B.2, we show that for every model with monotone choice behavior there exists a model with an independent threshold that has the same distribution of (M, Y^*, H, C) (Lemma A3).

two losses. If the belief is calibrated and the ratio carries no information about the outcome given the belief, this is a threshold model with H equal to their belief; in general, H also absorbs any part of their losses that is correlated with the outcome.

2.3 The Identified Set

Since consistency pins down $Q(M = m, Y^* = y) = P(M = m, Y^* = y)$, whether the decision maker's choices are consistent with a threshold model depends on the model only through two primitives: the distribution of their prediction given the algorithm's prediction and the outcome, and the probability that they select an individual with prediction h,

$$\rho_Q(h) := Q(C = 1 \mid H = h) = Q(T \geq h \mid H = h),$$

which we call the "choice probability function." The threshold rule imposes $\rho_Q(0) = 1$, since it selects every individual with prediction zero, and otherwise leaves the choice probability function unrestricted. Under the threshold model, the probability that the decision maker selects an individual given the algorithm's prediction and the outcome is then

$$Q(C = 1 \mid M = m, Y^* = y) = E_Q\big[\rho_Q(H) \mid M = m, Y^* = y\big]. \tag{1}$$

Equation (1) is bilinear in the two primitives: linear in each holding the other fixed.

Proposition 2.1. *The decision maker's choices are consistent with a threshold model $Q(\cdot)$ if and only if, for all $m \in \mathcal{M}$ and $y \in \{0, 1\}$,*

$$P(M = m, Y^* = y) = Q(M = m, Y^* = y), \tag{2}$$
$$P(C = 1 \mid M = m, Y^* = y) = E_Q\big[\rho_Q(H) \mid M = m, Y^* = y\big]. \tag{3}$$

Proposition 2.1 characterizes when the decision maker's choices are consistent with a given threshold model. The converse also holds: any calibrated distribution of (M, Y^*, H) and any function $\rho\colon [0, 1] \to [0, 1]$ with $\rho(0) = 1$ that together satisfy equations (2) and (3) are the primitives of some threshold model with which the decision maker's choices are consistent. The pairs of primitives of the threshold models in $\mathcal{Q}(P)$ are therefore exactly the pairs that satisfy equations (2) and (3).

Assumptions 1 and 2 enter Proposition 2.1 as restrictions on the choice probability function. Under monotone choice behavior, the choice probability function is nonincreasing, and under a fixed threshold t it is the step function $\rho_Q(h) = \mathbf{1}\{h \leq t\}$. We write $\mathcal{Q}^{\mathrm{MC}}(P)$ and $\mathcal{Q}^{\mathrm{FT}}(P)$ for the threshold models in $\mathcal{Q}(P)$ that satisfy each assumption, so that $\mathcal{Q}^{\mathrm{FT}}(P) \subseteq \mathcal{Q}^{\mathrm{MC}}(P) \subseteq \mathcal{Q}(P)$. Whether these sets are empty depends on the data through the outcome

rate among the individuals the decision maker selects, $\mu_1 := P(Y^* = 1 \mid C = 1)$, and the outcome rate among those they do not, $\mu_0 := P(Y^* = 1 \mid C = 0)$. In the pretrial example, these are the misconduct rate among released defendants and the misconduct rate among detained defendants.

Corollary 2.1. *i.* Monotone choice behavior. $\mathcal{Q}^{\mathrm{MC}}(P)$ *is nonempty if and only if* $\mu_1 < \mu_0$ *or* $C \perp\!\!\!\perp (M, Y^*)$ *under* $P(\cdot)$.

ii. Fixed threshold. $\mathcal{Q}^{\mathrm{FT}}(P)$ *is nonempty if and only if* $\mu_1 < \mu_0$ *under* $P(\cdot)$, *and the threshold must lie in the interval* $[\mu_1, \mu_0)$.

Corollary 2.1 shows that both assumptions are testable through a single comparison: the individuals the decision maker selects must be less risky on average than those they do not, $\mu_1 < \mu_0$. Under monotone choice behavior, the decision maker selects individuals less often the higher their prediction of an individual's risk, and so the individuals they select must have lower outcome rates on average (unless they select at random, in which case their choices are independent of the algorithm's prediction and the outcome). Under a fixed threshold, the decision maker sorts all individuals by their prediction and applies a single cutoff. This is the pooled analogue of the characterization in Rambachan (2024) (see also Caplin and Martin, 2015; Caplin and Dean, 2015).

2.4 Predictive Skill and Information Frontiers

Our first question compares the decision maker's predictive skill with the algorithm's: could choices based on their prediction alone outperform choices based on the algorithm's prediction? We next define predictive skill and the comparison formally under the threshold model.

A *signal* W is a function of (M, Y^*, H, T), such as the algorithm's prediction M, the decision maker's prediction H, their choice C, or the outcome Y^* itself. A *decision rule* $d\colon \mathcal{W} \to [0, 1]$ gives the probability of selecting an individual with signal value w,[5] and $\mathcal{D}(W)$ denotes the set of such decision rules. We evaluate a decision rule by its *score*, the pair $S(d; Q) := (Q(C^d = 1, Y^* = 1), Q(C^d = 1))$, where C^d is the choice induced by the decision rule: the fraction of individuals who are selected and have the outcome occur, and the fraction selected. Decision rules that lead to a lower first component and a higher second component are preferred.[6] In the pretrial example, the score is the fraction of defendants who are released and commit misconduct, and the release rate.

[5] A decision rule d is implemented as $C^d := \mathbf{1}\{U \leq d(W)\}$ for $U \sim \mathrm{Unif}[0, 1]$ independent of (M, Y^*, H, T).

[6] Appendix B.3 states the results for any pair of expected payoffs with the same orientation; with the false and true positive rates, for example, the information frontier is the receiver operating characteristic (ROC) curve.

Definition 2. The *feasible set* of a signal W under the threshold model $Q(\cdot)$ is the set of scores achievable by some decision rule based on it, $\mathcal{S}(W;Q) := \{S(d;Q) : d \in \mathcal{D}(W)\}$. The *information frontier* of W is the set of feasible scores that no other feasible score improves upon,

$$\mathcal{P}(W;Q) := \{s \in \mathcal{S}(W;Q) : \text{there is no } s' \in \mathcal{S}(W;Q) \text{ with } s'_1 \leq s_1,\ s'_2 \geq s_2,\ \text{and } s' \neq s\}.$$

The feasible set collects every score that rules based on W can deliver, and the information frontier is its Pareto frontier. We say the signal W *dominates* the signal W' if $\mathcal{S}(W';Q) \subseteq \mathcal{S}(W;Q)$.

Feasible sets and frontiers have several useful properties (Appendix B.3). The feasible set $\mathcal{S}(W;Q)$ is convex and compact. The feasible set contains the scores of selecting no one, $(0,0)$, and everyone, $(\mu,1)$, and the segment between them, the scores of selecting at random. The information frontier $\mathcal{P}(W;Q)$ is an increasing and concave curve that lies weakly above this segment. Any one-to-one transformation of W has the same feasible set, so the frontier measures the information the signal carries, not the scale on which it is recorded. Most importantly, the information frontier ranks signals exactly as Blackwell (1953) ranks experiments: W dominates W' if and only if W is more informative than W' about the outcome.

The decision maker's "predictive skill" is the information frontier of their prediction, $\mathcal{P}(H;Q)$, and the algorithm's is $\mathcal{P}(M;P)$.[7] In the pretrial release example, $\mathcal{P}(H;Q)$ is the set of best possible trade-offs between release-misconduct rates and release rates that are achievable using a judge's predictions, and $\mathcal{P}(M;P)$ is the set achievable using the algorithm's prediction. Comparing the predictive skill of the decision maker and the algorithm is equivalent to comparing these two information frontiers (see also Chan, Gentzkow and Yu, 2022).

3 Empirical Setting: New York Pretrial Release

We apply our framework to pretrial release decisions in New York State. Judges decide whether defendants are released before trial, and we construct an algorithm that predicts pretrial misconduct from the defendant's recorded characteristics. We apply our framework to empirically study three questions: what do judges' release decisions reveal about how well their information predicts pretrial misconduct compared with the algorithm? Would a designer who observes the algorithm's prediction gain by also eliciting a judge's release

[7] When an object depends on the threshold model only through the observable data, we write $P(\cdot)$ in place of $Q(\cdot)$; the algorithm's feasible set $\mathcal{S}(M;P)$, for example, is identified.

decision or their prediction? Would judges make better release decisions if they observed the algorithm's prediction?

3.1 Setting and Data

In New York State, a defendant who is arrested and charged appears before a judge at arraignment shortly after the arrest. The judge decides whether the defendant awaits trial in custody: the judge may release the defendant on recognizance or under non-monetary conditions such as supervised release, set monetary bail, or remand the defendant to custody. Since New York's bail reform took effect in January 2020, monetary bail and remand have been available only in cases charging a qualifying offense, and the statute directs the judge to the least restrictive conditions that reasonably assure the defendant's return to court.[8] The judge who conducts a defendant's arraignment is the judge staffing the arraignment calendar when the defendant is brought to court, determined by a rotating schedule rather than by the case (Kleinberg et al., 2017; Arnold, Dobbie and Hull, 2022). Because defendants are arraigned shortly after arrest and the arraignment rotation does not align with the assignment of prosecutors, defense attorneys, or trial judges, there is little scope to steer a case to a particular judge, and we treat assignment as quasi-random within a court, arraignment part and month. Section 3.3 presents supporting evidence.

We define release ($C = 1$) as release on recognizance, release under non-monetary conditions, or release on bail that is posted; a defendant whose bail is not posted or who is remanded is detained ($C = 0$). The outcome Y^* is whether the defendant would commit pretrial misconduct if released: either fail to appear at a required court date or be rearrested before the case is resolved, as in existing work on pretrial release (Kleinberg et al., 2017; Arnold, Dobbie and Hull, 2022; Angelova, Dobbie and Yang, 2023; Rambachan, 2024).[9]

We analyze records from the New York State Office of Court Administration that cover the universe of arraignments in New York State from 2020 to 2025. We restrict the sample to arraignments in which the judge makes a release decision that our framework can study: cases that are bail eligible, heard by an identifiable judge in a local court, not disposed at arraignment, and with a recorded release decision, as in Kleinberg et al. (2017) and Arnold, Dobbie and Hull (2022). We further restrict the sample to Black and white defendants charged with a felony or misdemeanor, to court-by-month strata with at least 100 cases

[8] N.Y. Crim. Proc. Law § 510.10(1). The statute allows the judge to consider only the defendant's return to court, not the risk to public safety. Amendments effective in July 2020 and May 2022 broadened the set of qualifying offenses, and amendments effective in June 2023 removed the requirement to select the least restrictive conditions (Rodriguez and Rempel, 2023).

[9] We assess the robustness of our results to narrower definitions of misconduct: whether the defendant would fail to appear alone, and whether the defendant would fail to appear or be rearrested for a violent felony.

and at least 2 judges, and to judges who decide at least 100 cases. The estimation sample contains 382,635 arraignments, decided by 371 judges across 13 courts and 723 court-by-month strata. Appendix D describes each restriction, and Appendix Table D.1 reports the cases that remain after each.

Table 1 summarizes release decisions and pretrial outcomes in the estimation sample. Judges release 79.6 percent of defendants, and 41 percent of released defendants commit pretrial misconduct. Judges differ substantially in their caseloads and release rates: the average judge decides 1,031 cases and releases 79.3 percent of them, with a standard deviation of 10.6 percentage points across judges. Release rates also differ across courts: the standard deviation of court-level release rates is 11.8 percentage points across the 13 courts. Misconduct rates among released defendants vary similarly: among the defendants the average judge releases, 40.4 percent commit pretrial misconduct, with a standard deviation of 7.3 percentage points across judges and 8.5 percentage points across courts.

Appendix Table D.2 reports defendant characteristics for all defendants, released defendants, and released defendants who commit misconduct. Detained defendants face more serious charges and have more extensive criminal histories than released defendants (84.9 percent of detained defendants face a felony charge, compared to 55.6 percent of released defendants), and within the released population a prior conviction or a pending case predicts misconduct.

Table 1: Release Decisions and Pretrial Misconduct across Courts and Judges

	Defendants (N = 382,635)	Courts (N = 13)		Judges (N = 371)	
		Mean	SD	Mean	SD
Number of cases		29,433	28,500	1,031	949
Release rate	0.796	0.737	0.118	0.793	0.106
No pretrial misconduct	0.592	0.601	0.085	0.596	0.073
Any pretrial misconduct	0.408	0.399	0.085	0.404	0.073
Failure to appear	0.195	0.212	0.064	0.193	0.064
Failure to appear or violent-felony rearrest	0.226	0.237	0.063	0.225	0.063
Other observed misconduct	0.182	0.162	0.044	0.179	0.042

Notes. This table reports release decisions and pretrial misconduct in the estimation sample of 382,635 arraignments in New York State between 2020 and 2025, decided by 371 judges in 13 courts. The first column reports shares of all arraignments. The remaining columns report the unweighted mean and standard deviation of each row across courts and across judges; the first row reports the number of arraignments per court and per judge. The release rate is the share of defendants released before trial, and the misconduct rows are shares of released defendants. Any pretrial misconduct is a failure to appear or any rearrest before the case is resolved. Failure to appear or violent-felony rearrest includes the failures to appear in the row above, and other observed misconduct is any rearrest not counted in the row above. See Section 3.1 for variable definitions and Appendix D for sample construction.

3.2 The Predictive Algorithm

We construct an algorithm that predicts pretrial misconduct from the characteristics recorded at arraignment: the defendant's demographics, the current charges, pending cases, criminal

history, and the court of arraignment (Appendix Table E.1). The algorithm is a gradient-boosted decision tree, trained on released defendants for whom misconduct is observed. We cross-fit the predictions: we split the estimation sample into two halves, train a model on each half, and score every case with the model trained on the other half, so every prediction is out of sample. The algorithm's predictions are further calibrated on held-out data before we use them. The cross-fitted, calibrated prediction is the algorithm's prediction M in our framework. Appendix E describes the estimation, tuning, and cross-fitting in more detail, and Appendix Figure E.2 summarizes performance. The area under the receiver operating characteristic curve (AUC) is 0.751, compared with 0.5 for a random guess and one for a perfect prediction; the AUC is the probability that the algorithm assigns a higher prediction to a randomly drawn released defendant who commits pretrial misconduct than to one who does not. Realized misconduct among released defendants rises from 7 percent in the lowest predicted-risk bin to 83 percent in the highest.

3.3 Selective Labels and Empirical Approach

Section 2 takes as given the joint distribution of the judge's choice, the algorithm's prediction, and the outcome, $(C, M, Y^*) \sim P(\cdot)$. In our data, there is a selective labels problem: misconduct is observed only for released defendants (Lakkaraju et al., 2017; Kleinberg et al., 2017). Writing observed misconduct as $Y := C \cdot Y^*$, the data identify the joint distribution of (C, M, Y). Consequently, the release-misconduct rate $p_{11}(m)$ and the misconduct rate among released defendants $\mu_1(m)$ are identified at every prediction m. The misconduct rate among detained defendants $\mu_0(m)$, and with it the outcome rate $\mu(m)$, is not directly identified.

We identify the outcome rates that our analysis requires using the quasi-random assignment of cases to judges, following the extrapolation approaches of Arnold, Dobbie and Hull (2022), Angelova, Dobbie and Yang (2023), and Baron et al. (2026). Because cases are quasi-randomly assigned to judges who differ in how many defendants they release, we can trace how release misconduct varies with the release rate and extrapolate to the outcome rate among all defendants. For intuition, a "supremely lenient" judge who releases every defendant would leave misconduct observed across their entire caseload, so their release-misconduct rate would identify the outcome rate $\mu = P(Y^* = 1)$ directly. No judge is that lenient, but judges differ in how many defendants they release, so we extrapolate across judges to a release rate of one. The same extrapolation on the cases at each prediction m identifies $\mu(m)$, and on each risk set $\{M \leq \tau\}$ it identifies the outcome rate among the defendants the algorithm would release at a cutoff τ.

Let Z denote the assigned judge and G the court-by-month stratum within which judges

are assigned, with w_g the share of cases in stratum g. We estimate the regressions

$$C_i = \sum_z \zeta_z \, \mathbf{1}\{Z_i = z\} + \sum_g \gamma_g^R \, \mathbf{1}\{G_i = g\} + u_i, \tag{4}$$

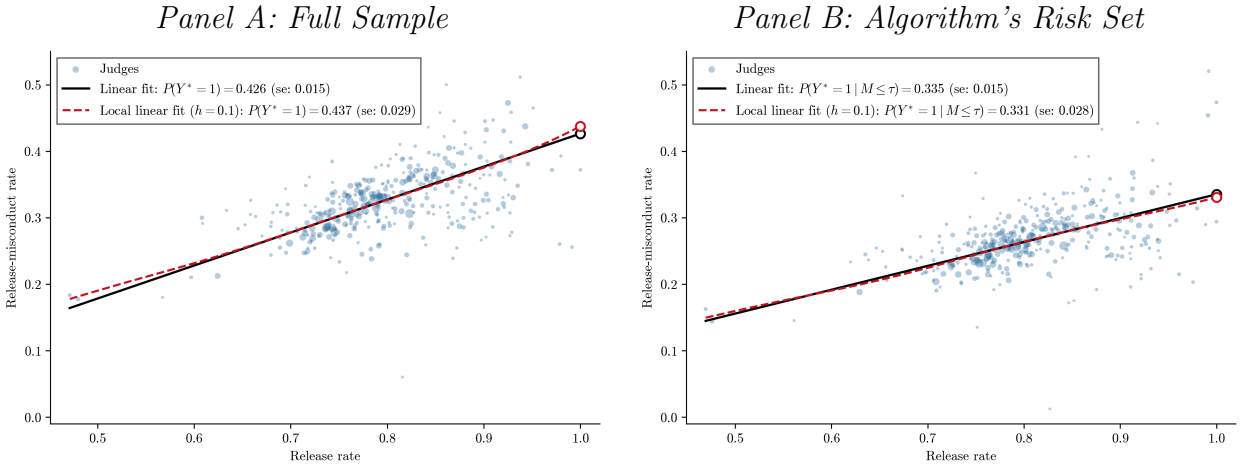
$$C_i Y_i = \sum_z \alpha_z \, \mathbf{1}\{Z_i = z\} + \sum_g \gamma_g^A \, \mathbf{1}\{G_i = g\} + v_i, \tag{5}$$

with one indicator for each judge, no intercept, and the first stratum omitted. Judge z's strata-adjusted release rate and release-misconduct rate are $\widehat{S}_{2z} := \widehat{\zeta}_z + \sum_g w_g \, \widehat{\gamma}_g^R$ and $\widehat{S}_{1z} := \widehat{\alpha}_z + \sum_g w_g \, \widehat{\gamma}_g^A$. We regress $\{\widehat{S}_{1z}\}_z$ on $\{\widehat{S}_{2z}\}_z$ across judges, weighting each judge by caseload, and the value of the fitted line at a release rate of one estimates μ. For the outcome rate at a prediction or on a risk set, we estimate the same regressions on the cases in that set and weight the stratum effects by that set's own shares w_g. We discretize the algorithm's prediction into deciles, so $m = 1, \ldots, 10$ indexes a decile of predicted risk. Standard errors come from a Bayesian bootstrap that draws weights for each case and recomputes every step within each replication (Rubin, 1981; Chamberlain and Imbens, 2003).

Figure 1 presents the resulting extrapolations for the full sample (Panel A) and for the risk set at the cutoff at which the algorithm releases as many defendants as the judges (Panel B). Each point is one judge's strata-adjusted release-misconduct rate against their strata-adjusted release rate, and the solid line is the caseload-weighted linear fit, whose value at a release rate of one is the extrapolated outcome rate. Release rates vary widely across judges and the most lenient judges release nearly all defendants, so the extrapolation is short. We estimate that 43 percent of defendants would commit misconduct if all were released (s.e. 1.5), and 34 percent (s.e. 1.5) among the defendants the algorithm would release at the judges' release rate. Appendix Figure F.2 presents the extrapolations within each decile of the prediction, where the fitted line passes through the origin because a judge who releases no defendant has no release misconduct; the estimated outcome rates rise from 10 percent in the bottom decile (s.e. 1.6) to 80 percent in the top (s.e. 0.4).

Our approach rests on three assumptions. First, cases are quasi-randomly assigned to judges within a court and month. Appendix Figure F.5 shows that case characteristics strongly predict the release decision (joint F-statistic of 827.3) but are much less predictive of the assigned judge's leave-out release rate (joint F-statistic of 3.8; within-strata R^2 of 0.005, compared to 0.124 for the release decision). Second, judges affect misconduct only through the release decision. Judges also set the conditions of release, such as monetary bail and supervision, and the restriction requires that these do not affect misconduct systematically across judges. This exclusion restriction is maintained throughout the literature on pretrial release in New York (Kleinberg et al., 2017; Arnold, Dobbie and Hull, 2022; Rambachan,

Figure 1: Identifying the Outcome Rate by Extrapolating Across Judges



Notes. This figure shows how we identify the outcome rate by extrapolating across quasi-randomly assigned judges. Each point is one judge's strata-adjusted release-misconduct rate (vertical axis) against their strata-adjusted release rate (horizontal axis), estimated from regressions (4) and (5). Panel A uses the full sample, and Panel B uses the risk set $\{M \leq \tau\}$ at the cutoff τ at which the algorithm releases as many defendants as the judges. In each panel, the solid line is the caseload-weighted linear fit and the dashed line a local linear fit. The open circles mark their values at a release rate of one, which estimate the outcome rate μ in Panel A and the outcome rate among the defendants the algorithm would release in Panel B; the legend reports each estimate. Standard errors are obtained from the Bayesian bootstrap (with 500 replications). See Section 3.3 for discussion.

2024). Third, the extrapolation recovers the limit it targets: the average relationship across judges between release-misconduct rates and release rates must be well approximated by the fitted line out to a release rate of one. The extrapolation is short because the most lenient judges release nearly every defendant, and a local linear fit gives 44 and 33 percent in the two panels of Figure 1. We do not require the first-stage monotonicity assumption that judges share a common ranking of defendants; the extrapolation requires only that the average relationship across judges be reliably estimated (see Arnold, Dobbie and Hull, 2022; Angelova, Dobbie and Yang, 2023).

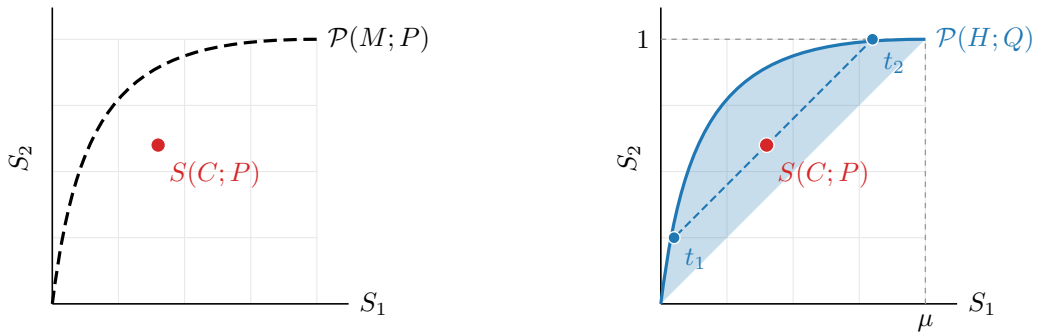
4 Comparing Humans and Algorithms

We now analyze how the decision maker's predictive skill compares with the algorithm's. Could choices based on the decision maker's predictions alone outperform choices based on the algorithm's prediction? We measure predictive skill by the information frontier of Section 2.4. The algorithm's frontier is identified from the data, but the decision maker's is not, as the data only record their choices. The common practice of comparing the decision maker's choices against the algorithm's frontier yields valid comparisons of predictive skill only under a strong restriction on their threshold, and we show how to bound their frontier instead.

4.1 Comparing Human Decisions with Algorithmic Predictions

The observable data identify the score of the decision maker's realized choices, which we call the *realized score* $S(C;P)$. A common empirical practice compares the realized score against the algorithm's frontier. If the realized score lies inside the algorithm's feasible set and off its frontier, then some decision rule based on the algorithm's prediction alone does weakly better on both components of the score and strictly better on one: the algorithm outperforms the decision maker's realized choices. Such comparisons appear in Kleinberg et al. (2017) for judges' release decisions, Mullainathan and Obermeyer (2021) for physicians' testing decisions, Angelova, Dobbie and Yang (2023) for judges' compliance with risk assessments, and Baron et al. (2026) for child protection screeners and investigators. Panel A of Figure 2 illustrates this common comparison.

Figure 2: Comparing Decisions with the Algorithm Can Understate the Decision Maker's Information



Panel A: Realized score against the algorithm's frontier

Panel B: Realized score against the decision maker's frontier

Notes. This figure illustrates Proposition 4.1. Scores are $S = (S_1, S_2)$, the selected-outcome rate and the selection rate. Panel A plots the decision maker's realized score $S(C;P)$ against the algorithm's information frontier $\mathcal{P}(M;P)$. Panel B plots the same realized score against the information frontier of their prediction, $\mathcal{P}(H;Q)$, under a threshold model whose threshold takes the two values t_1 and t_2: the shaded region is the feasible set of their prediction, the two marked points are the scores of the two cutoffs, and their choices mix the two, so the realized score lies strictly inside their frontier. See Section 4.1 for discussion.

However, we wish to study the decision maker's information frontier: could any rule based on their prediction improve on the algorithm's prediction? Their choices apply their threshold to their prediction, so the quality of their choices and the quality of their prediction can differ. We characterize when the common comparison of the decision maker's choices versus the algorithm's predictions yields valid comparisons of predictive skill.

Proposition 4.1. *Suppose the decision maker's choices are consistent with a threshold model $Q(\cdot)$. The realized score lies on their information frontier if and only if there is a cutoff $\tau \in [0,1]$ such that $C = \mathbf{1}\{H \leq \tau\}$ almost surely on the event $\{H \neq \tau\}$.*

Proposition 4.1 states that the realized score lies on their information frontier if and only if their choices are those of a fixed threshold applied to their prediction, up to ties. The common comparison therefore measures their predictive skill only if they apply a single cutoff to their prediction, up to ties: it assumes away any heterogeneity in their preferences and noise. Otherwise, the realized score lies strictly inside their information frontier, and the gap between their realized score and the algorithm's frontier mixes their predictive skill with the variation in their threshold. The intuition is Jensen's inequality. For simplicity, suppose the decision maker's threshold is drawn independently of their prediction and takes two values, $t_1 < t_2$. Each value yields a cutoff rule: at a given selection rate, it selects the individuals with the lowest outcome rates, and so produces a point on their information frontier. Their realized choices mix the two rules, applying t_1 to some individuals and t_2 to the others, so the realized score is an average of two distinct frontier points, weighted by the probability of each threshold value. Since the information frontier is concave, the average lies strictly inside it, unless the individuals the two values treat differently all share a single value of their prediction, as illustrated in Panel B of Figure 2. The same logic applies to any threshold that takes more than one value, and Proposition 4.1 shows that their realized score lies on their frontier only if the threshold varies without changing any choice, up to ties.

4.2 Bounding the Information Frontier under the Threshold Model

Under the threshold model, what do the data reveal about the decision maker's information frontier in general? As $Q(\cdot)$ ranges over the identified set $\mathcal{Q}(P)$, the feasible set of their prediction $\mathcal{S}(H;Q)$ ranges over a family of sets. The *upper bound* is the smallest closed set containing every member of the family, and the *lower bound* is the intersection of its members:

$$\overline{\mathcal{S}}(P) := \overline{\bigcup_{Q\in\mathcal{Q}(P)} \mathcal{S}(H;Q)}, \qquad \underline{\mathcal{S}}(P) := \bigcap_{Q\in\mathcal{Q}(P)} \mathcal{S}(H;Q).$$

A score lies in the upper bound if some consistent threshold model justifies it: under that model, a rule based on their prediction attains the score or comes arbitrarily close. A score lies in the lower bound if every consistent threshold model justifies it. The upper bound is therefore what the data cannot rule out about their predictive skill, and the lower bound is what the data guarantee. Substituting $\mathcal{Q}^{\mathrm{MC}}(P)$ or $\mathcal{Q}^{\mathrm{FT}}(P)$ for $\mathcal{Q}(P)$ in the display above gives the bounds under monotone choice behavior, $\overline{\mathcal{S}}^{\mathrm{MC}}(P)$ and $\underline{\mathcal{S}}^{\mathrm{MC}}(P)$, and under a fixed threshold, $\overline{\mathcal{S}}^{\mathrm{FT}}(P)$ and $\underline{\mathcal{S}}^{\mathrm{FT}}(P)$, respectively.

Bounds on the feasible set are bounds on the information frontier, because a larger feasible set has a weakly higher frontier. They permit comparisons of predictive skill that

hold under every consistent threshold model: the algorithm dominates the decision maker's prediction under every such model if and only if $\overline{\mathcal{S}}(P) \subseteq \mathcal{S}(M;P)$, and their prediction dominates the algorithm under every such model if and only if $\mathcal{S}(M;P) \subseteq \underline{\mathcal{S}}(P)$.

In analyzing what the data reveal about the decision maker's information frontier, we introduce three feasible sets as references. We say that a signal W *reveals the outcome* under a threshold model $Q(\cdot)$ if $Q(Y^* = 1 \mid W) \in \{0, 1\}$ almost surely. The *perfect-prediction set* Π is the feasible set of a signal that reveals the outcome, the largest feasible set any signal can have: its frontier runs from $(0, 0)$ to $(0, 1-\mu)$ and then to $(\mu, 1)$. The *random-selection set* $\mathcal{S}(\varnothing;P)$ is the feasible set of an uninformative signal, the segment from $(0, 0)$ to $(\mu, 1)$. Every feasible set contains the random-selection set and lies in the perfect-prediction set. Finally, the decision maker's choice C is itself a signal, and its feasible set $\mathcal{S}(C;P)$ is identified: its frontier runs from $(0, 0)$ through the realized score to $(\mu, 1)$, the scores of rules that follow, reverse, or randomize over their choice. Appendix Figure A1 visualizes these three feasible sets.

Whenever we impose monotone choice behavior or a fixed threshold, we assume that some consistent threshold model satisfies it (Corollary 2.1). We say that there is *no reversal across the algorithm's predictions* if

$$\max_{m \in \mathcal{M}} P(C = 1 \mid M = m, Y^* = 1) \leq \min_{m \in \mathcal{M}} P(C = 1 \mid M = m, Y^* = 0),$$

that is, if individuals for whom the outcome occurs are selected, at every prediction of the algorithm, no more often than individuals for whom it does not are selected at any prediction.

Proposition 4.2. *Under the threshold model, the upper bound on the feasible set of the decision maker's prediction is the perfect-prediction set,* $\overline{\mathcal{S}}(P) = \Pi$*, and the lower bound is the feasible set of their own choices,* $\underline{\mathcal{S}}(P) = \mathcal{S}(C;P)$*.*

i. Monotone choice behavior. *The upper bound* $\overline{\mathcal{S}}^{\mathrm{MC}}(P)$ *is* Π *if and only if there is no reversal across the algorithm's predictions; otherwise it lies strictly inside* Π*.*

ii. Fixed threshold. *The upper bound* $\overline{\mathcal{S}}^{\mathrm{FT}}(P)$ *lies strictly inside* Π*.*

In both cases, the lower bound remains $\mathcal{S}(C;P)$*.*

Consider first the lower bound, which is what the data guarantee about their predictive skill. Under the threshold model, the decision maker's choice is a garbling of their prediction, and their prediction can deliver every score that their choice can. The lower bound on the feasible set of their prediction is therefore exactly the feasible set of their own choices, under the threshold model alone and under either restriction: the data can never guarantee that their prediction supports better decisions than their own choices.

Consider next the upper bound, which is what the data cannot rule out about the decision maker's information frontier. Under the threshold model alone, it is the perfect-prediction set for every distribution of the data. The threshold model leaves the distribution of their threshold given their prediction unrestricted, so their choice probability function may rise with their prediction. A prediction that nearly reveals the outcome places the individuals with $Y^* = 1$ near one and those with $Y^* = 0$ near zero, and nothing prevents the decision maker from selecting the former more often than the latter. Any choices are therefore consistent with a nearly revealing prediction, and the algorithm can never be shown to dominate it.

Monotone choice behavior removes this freedom: their choice probability function must be nonincreasing. With a nearly revealing prediction, the decision maker must then select the individuals with $Y^* = 0$ at least as often as those with $Y^* = 1$, across the algorithm's predictions as well as within each, so there is no reversal across the algorithm's predictions. When that condition holds, there exists a threshold model satisfying monotone choice behavior with a nearly revealing prediction. When it fails, the decision maker selects the individuals with $Y^* = 1$ at some prediction of the algorithm more often than those with $Y^* = 0$ at another. Consequently, the decision maker's prediction must pool individuals with and without the outcome, the upper bound lies strictly inside Π, and the data may now show that the algorithm dominates the decision maker's prediction.

Under a fixed threshold, the upper bound lies strictly inside Π for every distribution of the data. If their prediction nearly revealed the outcome, a fixed threshold would select nearly every individual with $Y^* = 0$ and nearly none with $Y^* = 1$ at every prediction of the algorithm, and the data rule this out. By Proposition 4.1, the realized score lies on the frontier of their prediction, so at the decision maker's own selection rate the upper and lower bounds coincide. At other selection rates, the bounds differ. A fixed threshold determines which individuals the decision maker selects, but the data do not reveal how their prediction varies among the individuals they select or among those they do not. Their prediction may, for example, equal zero for some of the individuals they select and equal the threshold for the rest. A lower cutoff would then select only individuals with $Y^* = 0$, whereas any rule based on their choice alone selects a random subset of those they select. The upper bound is the union of the feasible sets of two calibrated predictions that each take two values (Appendix A.2).

We report the bounds on the information frontier for a few judges, but bounds on a set are difficult to summarize across many decision makers. We therefore summarize the information frontier by focusing on its value at a single selection rate. For a signal W and a selection rate $r \in [0, 1]$, the frontier of W at rate r is the lowest selected-outcome rate that

a rule based on W attains while selecting a fraction r of individuals,

$$\tilde{S}_W(r;Q) := \min\big\{S_1(d;Q) : d \in \mathcal{D}(W),\ S_2(d;Q) = r\big\}, \tag{6}$$

the point on the information frontier of W at selection rate r. The algorithm's frontier at rate r, $\tilde{S}_M(r;P)$, is identified. The bounds on the decision maker's frontier at rate r are

$$\underline{\tilde{S}}_H(r;P) := \inf_{Q\in\mathcal{Q}(P)} \tilde{S}_H(r;Q) \qquad \text{and} \qquad \overline{\tilde{S}}_H(r;P) := \sup_{Q\in\mathcal{Q}(P)} \tilde{S}_H(r;Q),$$

the points at rate r on the frontiers of $\overline{\mathcal{S}}(P)$ and $\underline{\mathcal{S}}(P)$, with the superscripts MC and FT when the identified set is restricted. The first is the best that rules based on their prediction could do at rate r under some consistent threshold model, and the second is the best they are guaranteed to do under every such model. Each bound is the optimal value of a bilinear program in the primitives of Section 2.3 (Appendix C.1).

4.3 Comparing Judges and the Algorithm in New York

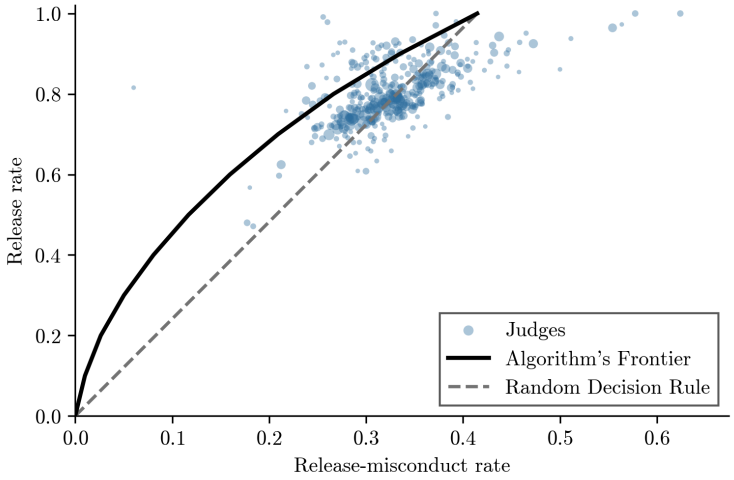
With our empirical approach described in Section 3.3, the joint distribution $(C, M, Y^*) \sim P(\cdot)$ is identified, and we can estimate both comparisons of the judges against the algorithm in New York pretrial release.

4.3.1 Comparing Judges' Decisions with Algorithmic Predictions

Figure 3 presents the common comparison for the judges in our sample, the empirical counterpart of Panel A of Figure 2. The algorithm's frontier is traced from the extrapolated outcome rates of Section 3.3: at each cutoff τ, the rule that releases every defendant with $M \leq \tau$ has a release rate of $P(M \leq \tau)$ and a release-misconduct rate of $P(Y^* = 1, M \leq \tau)$, the extrapolated outcome rate on the risk set multiplied by its share of defendants, and varying the cutoff over a grid traces the frontier. Each point is one judge's strata-adjusted release-misconduct rate against their strata-adjusted release rate, sized by caseload, and the random-selection line runs from the origin to $(\mu, 1)$, with μ the extrapolated outcome rate for all defendants.

Judges vary widely along both axes. Release rates range from 47 to 100 percent, and judges with similar release rates release defendants of very different risk. Most judges lie inside the algorithm's frontier: of the 371 judges, 323 (87%) do, and they decide 91% of cases. Because each judge's realized score is estimated with sampling error, we compute empirical Bayes estimates of the share of judges whose true scores lie inside the algorithm's frontier, using the distribution of the judges' point estimates and their sampling error (following

Figure 3: Comparing Judges' Decisions with the Algorithm's Predictions in New York



Notes. This figure reports the common comparison for the 371 judges in the estimation sample. Each point is one judge's strata-adjusted release-misconduct rate (horizontal axis) against their strata-adjusted release rate (vertical axis), estimated as in Section 3.3, with points sized by the judge's caseload. The solid line is the algorithm's frontier, traced from the extrapolated outcome rates of the risk sets $\{M \leq \tau\}$ over a grid of cutoffs τ, and the dashed line is the random decision rule, the line from the origin to $(\mu, 1)$, with μ the extrapolated outcome rate for all defendants. See Section 4.3.1 for discussion.

Bonhomme and Weidner, 2021). We estimate that 95 percent of judges' scores lie inside the algorithm's frontier: at their own release rates, the algorithm would release the same share of defendants with less misconduct. The gap is large in the aggregate: at the judges' overall release rate, the defendants the algorithm would release commit misconduct at an extrapolated rate of 34 percent (s.e. 1.5), compared to 41 percent among the defendants the judges release.

By Proposition 4.1, a judge's realized score lies on the frontier of their prediction only if they apply a fixed threshold, up to ties, so the gap between a judge's score and the algorithm's frontier measures their predictive skill only under that restriction. The restriction is empirically testable. By Corollary 2.1, a fixed threshold is consistent with a judge's choices if and only if the defendants they release are less likely to commit misconduct than those they detain, that is, if and only if their realized score lies above the random-selection line; monotone choice behavior requires the same unless the judge's choices are independent of the prediction and the outcome. Figure 3 includes the line, and many judges lie below it: 47 percent by their point estimates (176 judges) and 43 percent by the empirical Bayes estimate. For these judges, the data reject the preference restriction under which the common comparison of human decisions versus algorithmic predictions would measure predictive skill. For the others, the figure is silent. A judge whose threshold varies has a realized score strictly inside their own frontier, and they may outperform the algorithm even though their score lies

inside the algorithm's frontier. We therefore bound each judge's information frontier under the threshold model alone, under monotone choice behavior, and under a fixed threshold next.

4.3.2 Bounding the Judges' Information Frontiers

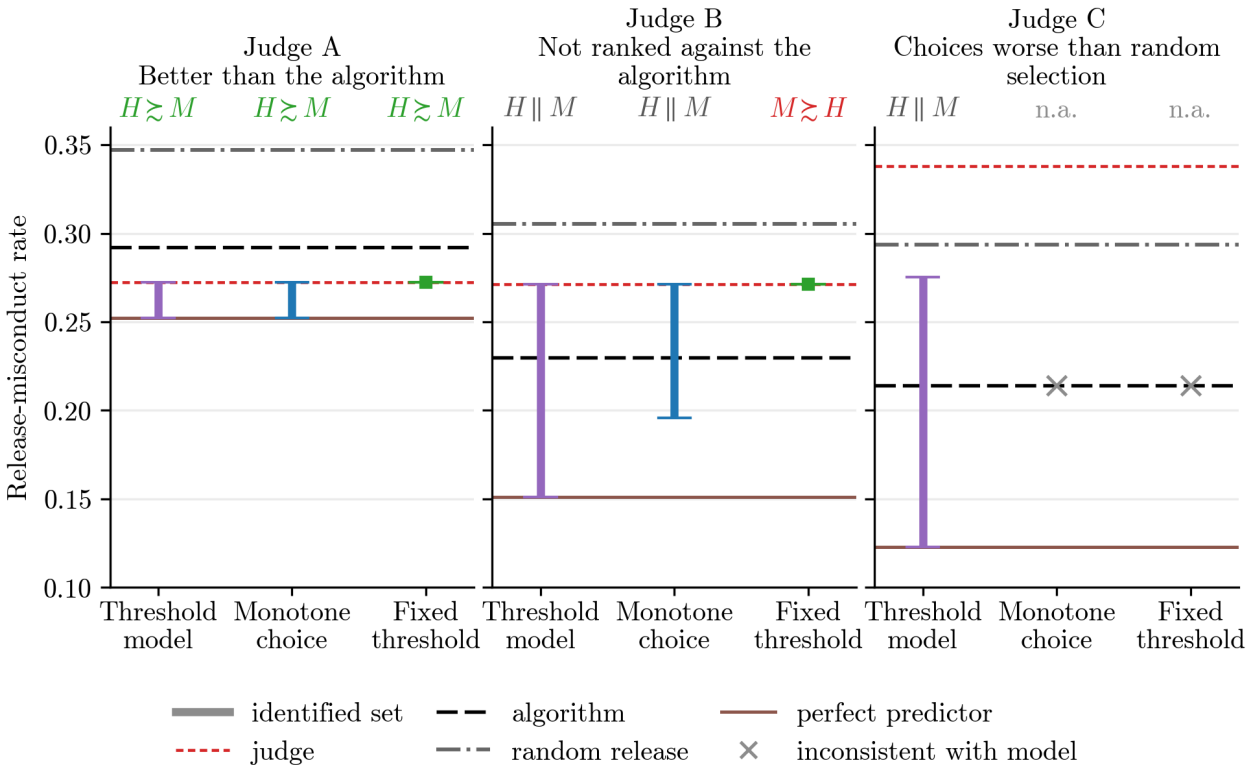
We compute the bounds on each judge's information frontier at the judge's own release rate using the fixed-rate bounds discussed in Section 4.2. Among decision rules that release the same share of defendants as the judge, the lower bound is the lowest release-misconduct rate that the judge's prediction could support under some consistent threshold model, and the upper bound is the lowest it is guaranteed to support under every such model. The bounds therefore summarize the release-misconduct rates a designer could achieve using the judge's prediction while holding the judge's release rate fixed.

By Proposition 2.1, a threshold model consistent with a judge's choices is described by the distribution of their prediction given the algorithm's prediction and the outcome and by their choice probability function, and the lowest and highest release-misconduct rates that their prediction could deliver at their release rate are the optimal values of two programs in these unknowns, whose constraints, equations (2) and (3), are bilinear. We solve the programs judge by judge, replacing the population quantities in the constraints by their empirical counterparts: the judge's release rate and release-misconduct rate within each decile of predicted risk, strata-adjusted at the decile's own mix of strata as in Section 3.3, and the extrapolated outcome rate of each decile. Appendix C.1 provides more detail.

Figure 4 presents the bounds for three judges, chosen by where their realized scores fall in Figure 3: Judge A's lies above the algorithm's frontier, Judge B's between the frontier and the random-selection line, and Judge C's below the line. Judge A releases 84 percent of defendants, more than Judges B and C at 74 and 71 percent, and 33 percent of the defendants Judge A releases commit misconduct, against 37 and 48 percent. For each judge, the bars give the range of release-misconduct rates that the judge's prediction could produce at the judge's own release rate as restrictions are added, alongside the algorithm's rate at that release rate, random release, and the perfect predictor. Judges A and B are consistent with the threshold model and both restrictions. Under the threshold model alone and under monotone choice behavior, Judge A could be as good as the perfect predictor; under a fixed threshold, the bounds collapse to Judge A's realized rate (Proposition 4.1), which still lies below the algorithm's, so Judge A beats the algorithm under every restriction. Judge B could also match the perfect predictor under the threshold model alone; monotone choice behavior rules this out but leaves Judge B's best case below the algorithm's rate, so the comparison is inconclusive, and only under a fixed threshold does the algorithm beat Judge

B. Judge C's choices are worse than random release, so neither restriction is consistent with them (Corollary 2.1). Under the threshold model alone, Judge C could still be as good as the perfect predictor, because the threshold model lets the threshold vary with the prediction, and a judge whose prediction reveals the outcome can then select worse than random in the aggregate.

Figure 4: Bounds on Three Judges' Information Frontiers at Their Release Rates



Notes. This figure reports the bounds on the release-misconduct rate that three judges' predictions could support at each judge's own release rate. Each panel is one judge; the panel title gives the judge's position in Figure 3 and their release rate. Each bar runs from the lower to the upper bound of Section 4.2 under the restrictions on the horizontal axis, cumulative from left to right (Proposition 4.2), and a cross marks a restriction that is inconsistent with the judge's choices (Corollary 2.1). The labels above the bars compare the judge with the algorithm at the judge's release rate: $H \gtrsim M$, the judge's prediction supports a lower release-misconduct rate than the algorithm under every consistent model; $M \gtrsim H$, the algorithm does; $H \parallel M$, neither. The solid brown line is the perfect predictor, the black dashed line the algorithm, the gray dash-dot line random release, and the red dashed line the judge's realized rate. Deciles are weighted equally, and the bars are point estimates. See Section 4.3.2 for discussion.

Figure 5 traces the bounds across release rates, which visualizes the bounds on each judge's full information frontier. Each panel shows the region between the upper and lower bounds on a judge's frontier under one combination of restrictions, together with the algorithm's frontier and random selection. The common comparison speaks only to the release rate the judge chose, and so do the bars in Figure 4. The regions in Figure 5 bound what

the judge's prediction could achieve at every other release rate, for instance if the judge were required to release fewer defendants. Under the threshold model alone, the upper bound for every judge is the frontier of the perfect predictor, and the data say little at any release rate. Monotone choice behavior pulls the upper bounds of Judges A and B below the perfect predictor at every release rate, not only at their own. A fixed threshold narrows the region further around the frontier of the judge's own choices, and for Judge B the region lies close to the algorithm's frontier at every release rate. Under either restriction, the data therefore bound a judge's predictive skill at release rates far from the one the judge actually chose.

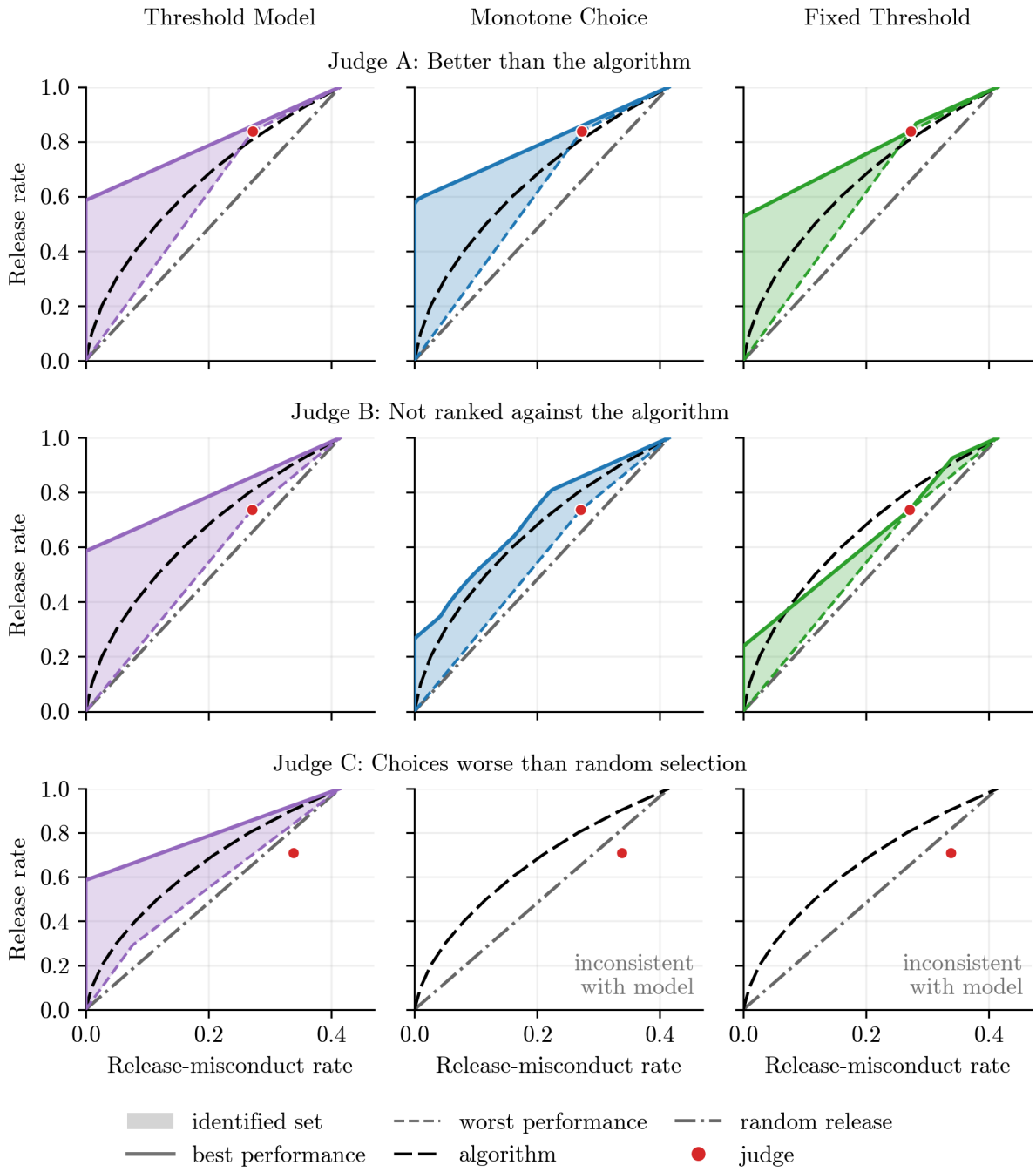
Judge C's choices are worse than random, yet the lower bound on Judge C's information frontier lies above the judge's realized score. This is because worse-than-random choices still carry information about misconduct, albeit in the wrong direction. A rule that releases the defendants Judge C detains uses this information better. For Judges A and B, whose realized scores lie above the random-selection line, the lower bound instead connects the origin, the realized score, and full release.

Table 2 reports the bounds at each judge's own release rate for the 362 judges who hear at least one case in every decile of predicted risk, which the identified set requires.

Every judge is consistent with the threshold model, but only 61.6 percent (90 percent credible interval from 58.8 to 76.5) are consistent with monotone choice behavior or a fixed threshold; the two shares coincide because no judge's choices are independent of the prediction and the outcome (Corollary 2.1(i)). The remaining judges release riskier defendants than they detain, so no threshold model with either restriction rationalizes their choices, and the gap between their realized score and the algorithm's frontier need not reflect a gap in predictive skill. The data alone cannot say why these judges do so: for example, their thresholds may vary with the charge, the statute directs them to the defendant's return to court rather than the outcome we measure, or they may use their prediction in the wrong direction.

Panel A reports the share of judges consistent with each combination of restrictions and, among them, how the judge's prediction compares with the algorithm at the judge's release rate: the judge beats the algorithm if the upper bound on their release-misconduct rate lies below the algorithm's, the algorithm beats the judge if the lower bound lies above it, and otherwise the two cannot be ranked. How the judges compare depends heavily on which restriction we impose. Under the threshold model alone, 99.4 percent of judges cannot be ranked. Monotone choice behavior alone ranks more than half of judges, and a fixed threshold ranks every judge, with the algorithm beating 99.1 percent. The shares also show how typical the three judges in Figure 4 are. Judges like A, who beat the algorithm under every restriction, are 0.6 percent of judges; judges like B, who cannot be ranked under

Figure 5: Bounds on Three Judges' Information Frontiers



Notes. This figure reports the bounds on three judges' information frontiers at every release rate. Each row is one judge, with the judge's position in Figure 3 in the row title, and each column is one combination of restrictions, cumulative from left to right. The shaded region runs between the frontiers of the upper and lower bounds on the feasible set of the judge's prediction (Proposition 4.2), computed with the programs of Appendix C.1 at release rates in steps of 0.01; a panel marked inconsistent is a restriction that the judge's choices reject (Corollary 2.1). The black dashed curve is the algorithm's information frontier, the gray dash-dot line the random-selection line, and the red point the judge's realized score. Deciles are weighted equally, and the bounds are point estimates. See Section 4.3.2 for discussion.

monotone choice behavior but whom the algorithm beats under a fixed threshold, are 27.3 percent; judges like C, consistent with the threshold model alone, are 38.4 percent; and the algorithm beats the remaining 33.7 percent under monotone choice behavior.

Panel B reports the size of these comparisons: the bounds on the judges' release-misconduct rate minus the algorithm's and minus that of random release, averaged across the judges consistent with each combination of restrictions. Under the threshold model alone, the judges' rate could lie anywhere from 5.9 percentage points below to 5.2 above the algorithm's. Under monotone choice behavior it lies between 0.4 and 4.7 percentage points above the algorithm's, and under a fixed threshold it is 4.7 points above; under all three, the judges' prediction supports a lower rate than random release. Monotone choice behavior alone therefore narrows the comparison substantially without a fixed threshold, and most of what the fixed threshold adds is to rank the judges like B.

Table 2: Bounds on the Judges' Information Frontiers at Their Own Release Rates

	Model restrictions		
	(1) Threshold Model	(2) Monotone Choice	(3) Fixed Threshold
Panel A: comparison at the observed release rate			
Judges consistent with the model	100.0	61.6	61.6
	[100.0, 100.0]	[58.8, 76.5]	[58.8, 76.5]
Human $\succsim$ Algorithm	0.6	0.9	0.9
	[0.8, 3.0]	[1.2, 4.4]	[1.2, 4.4]
Human $\parallel$ Algorithm	99.4	44.4	0.0
	[97.0, 99.2]	[33.2, 47.0]	[0.0, 0.0]
Algorithm $\succsim$ Human	0.0	54.7	99.1
	[0.0, 0.0]	[49.3, 64.4]	[95.6, 98.8]
Panel B: release-misconduct rate compared with benchmarks			
Human minus algorithm			
Upper bound	5.2	4.7	4.7
	[4.5, 5.2]	[4.0, 4.7]	[4.0, 4.7]
Lower bound	−5.9	0.4	4.7
	[−6.3, −5.9]	[0.1, 1.1]	[4.0, 4.7]
Human minus random release			
Upper bound	−1.3	−1.9	−1.9
	[−2.0, −1.4]	[−2.5, −1.9]	[−2.5, −1.9]
Lower bound	−12.4	−6.1	−1.9
	[−12.8, −12.4]	[−6.4, −5.4]	[−2.5, −1.9]

Notes. This table reports the bounds on the release-misconduct rate that each judge's prediction could support at the judge's own release rate, for the 362 judges who hear at least one case in each decile of predicted risk, with equal decile weights. Panel A reports the percent of these judges consistent with each combination of restrictions and, as a percent of the judges consistent with it, the comparison with the algorithm at the judge's release rate: Human $\succsim$ Algorithm when the upper bound on the release-misconduct rate of the judge's prediction lies below the algorithm's, Algorithm $\succsim$ Human when the lower bound lies above it, and Human $\parallel$ Algorithm otherwise. Panel B reports the release-misconduct rate of the judge's prediction minus that of the algorithm and minus that of random release, averaged across the judges consistent with each combination, in percentage points; negative values favor the judge. Brackets are 90 percent credible intervals calculated over 500 Bayesian-bootstrap draws. See Section 4.3.2 for discussion.

5 The Value of Human Input

We now turn to the second question: would a designer who observes the algorithm's prediction gain by also eliciting the decision maker's choice or their prediction? The data identify whether the decision maker has private information and the value of eliciting their choice. The value of eliciting their prediction is only partially identified under the threshold model, and we bound it under each restriction on their threshold. We then test for private information and bound the value of eliciting the judges' input in New York pretrial release.

5.1 Detecting Private Information

The value of eliciting the decision maker's input depends first on whether they have anything to add: do they know something about the outcome beyond the algorithm's prediction?

Definition 3. The decision maker *has private information at prediction* $m \in \mathcal{M}$ if $H \not\perp\!\!\!\perp Y^* \mid M = m$ under every threshold model $Q(\cdot)$ consistent with their choices. They *have private information* if they have private information at some $m \in \mathcal{M}$.

That is, among individuals with the same prediction of the algorithm, the decision maker's prediction adds signal for the outcome, under every model that rationalizes their choices. Private information may come from recorded characteristics that the algorithm did not use, from features that are never recorded, or from any other factor in their choices that is correlated with the outcome (Section 2.2). The definition is stated prediction by prediction: they may have private information for some individuals and not for others.

Private information is equivalent to a statement about the observable data, and to a statement about the value of the decision maker's input. For a weight $\lambda > 0$ on the two components of the score, the *value* of a signal W is the largest expected payoff that a decision rule based on it can achieve,

$$V_\lambda(W; Q) := \max_{d \in \mathcal{D}(W)} S_2(d; Q) - \lambda\, S_1(d; Q).$$

The values of the algorithm's prediction and of the algorithm's prediction together with the decision maker's choice are identified and denoted $V_\lambda(M; P)$ and $V_\lambda((M, C); P)$, respectively.

Lemma 2.

i. The decision maker has private information at prediction $m \in \mathcal{M}$ if and only if $C \not\perp\!\!\!\perp Y^* \mid M = m$.

ii. The decision maker has private information if and only if $V_\lambda((M, C); P) > V_\lambda(M; P)$ for some $\lambda > 0$.

Lemma 2(i) turns a statement about every model in the identified set into a statement about the data: if the decision maker's choice and the outcome are dependent at prediction m, no threshold model rationalizes their choices unless their prediction is informative about the outcome. Otherwise, there is some threshold model that rationalizes their choices with a prediction that is independent of the outcome. Because the choice and the outcome are binary, $C \not\perp\!\!\!\perp Y^* \mid M = m$ is equivalent to $\mu_1(m) \neq \mu_0(m)$: testing for private information at prediction m compares the outcome rate among the individuals they select with the rate among those they do not (Appendix Figure A2). Lemma 2(ii) restates the test as a statement about the value of their input: they have private information exactly when a designer who holds the algorithm's prediction could do strictly better, at some weight λ, by also observing their choice.

Evidence of private information is robust: a dependence between the choice and the outcome at prediction m cannot be rationalized without an informative prediction. The absence of such evidence is not evidence that the decision maker's prediction is uninformative. Their choice is a garbling of their prediction: it reveals only whether their prediction lies above or below their threshold, and their prediction may still vary among the individuals they select at prediction m and among those they do not. If the choice is informative about the outcome at m, their prediction must be, but the converse fails: because calibration holds pooled across the algorithm's predictions rather than within each of them, there are calibrated threshold models that change none of their choices under which H is informative at m, although the choice is not. Closing this gap requires a restriction on how their prediction relates to the outcome within each prediction of the algorithm, which we discuss in Section 7.

The same characterization holds under either restriction on their threshold: whenever some consistent threshold model satisfies monotone choice behavior or a fixed threshold, the decision maker has private information at m under every such model if and only if $C \not\perp\!\!\!\perp Y^* \mid M = m$.

5.2 Bounding the Value of Human Input under the Threshold Model

Consider a designer who must decide which individuals to select, has the algorithm's prediction, and can ask the decision maker for input. The designer's payoff from a decision rule d is $S_2(d;Q) - \lambda\, S_1(d;Q)$ for some weight $\lambda > 0$: in the pretrial example, the designer trades off a higher release rate against a higher release-misconduct rate. The designer can base decisions on the algorithm's prediction alone, on the algorithm's prediction together with the decision maker's choice, or on the algorithm's prediction together with the decision

maker's prediction, with values $V_\lambda(M;P)$, $V_\lambda((M,C);P)$, and $V_\lambda((M,H);Q)$, respectively.[10]

By Lemma A5, the optimal rule based on a signal selects the individuals whose outcome probability given the signal lies below the cutoff $1/\lambda$. For a signal with finite support, its value is $V_\lambda(W;Q) = \sum_w \big(Q(W=w) - \lambda\, Q(W=w, Y^*=1)\big)_+$. Based on the algorithm's prediction alone, the rule selects every individual at the predictions with $\mu(m) < 1/\lambda$. Eliciting the decision maker's choice refines the predictions: given (M,C), a selected individual at prediction m has outcome probability $\mu_1(m)$ and an unselected one $\mu_0(m)$, so the rule applies the cutoff at each prediction. The gain from eliciting the choice, $V_\lambda((M,C);P) - V_\lambda(M;P)$, is strictly positive if and only if the two rates bracket the cutoff at some prediction, so private information is necessary for the choice to be worth eliciting but not sufficient.

The choice compresses the decision maker's prediction into one bit, $C = \mathbf{1}\{H \le T\}$: it mixes their prediction with their threshold, and it discards the variation in their prediction on each side of the threshold. Eliciting their prediction removes both losses. Under the threshold model, their prediction summarizes everything in their choices that is correlated with the outcome, so $V_\lambda((M,H);Q)$ provides a ceiling on what any mechanism for eliciting their input could deliver. Since $V_\lambda((M,H);Q)$ depends on the threshold model, it is only partially identified. As $Q(\cdot)$ ranges over the identified set $\mathcal{Q}(P)$, its *upper bound* $\overline{V}_\lambda(P)$ is its supremum, the best value the data cannot rule out, and its *lower bound* $\underline{V}_\lambda(P)$ is its infimum, what the data guarantee. Substituting $\mathcal{Q}^{\mathrm{MC}}(P)$ or $\mathcal{Q}^{\mathrm{FT}}(P)$ gives the bounds under monotone choice behavior, $\overline{V}^{\mathrm{MC}}_\lambda(P)$ and $\underline{V}^{\mathrm{MC}}_\lambda(P)$, and under a fixed threshold, $\overline{V}^{\mathrm{FT}}_\lambda(P)$ and $\underline{V}^{\mathrm{FT}}_\lambda(P)$. At a weight $\lambda \le 1$ the cutoff is at least one, every signal has the same value $1-\lambda\mu$, and we therefore consider weights $\lambda > 1$.

The bounds are characterized by optimization: each is the optimal value of a bilinear program in the primitives of Section 2.3 (Appendix C.2). When we report the bounds at a fixed selection rate, as in Section 4.2, the weight is replaced by a constraint on the selection rate: the frontier of the pair at rate r, $\tilde{S}_{(M,H)}(r;Q)$ in equation (6), has the bounds $\underline{\tilde{S}}_{(M,H)}(r;P) := \inf_{Q\in\mathcal{Q}(P)} \tilde{S}_{(M,H)}(r;Q)$ and $\overline{\tilde{S}}_{(M,H)}(r;P) := \sup_{Q\in\mathcal{Q}(P)} \tilde{S}_{(M,H)}(r;Q)$, with the superscripts MC and FT under the restrictions. Appendix Figure A3 illustrates both measures: at a weight λ, the value of the decision maker's prediction to the designer is the vertical distance between the lines of slope λ tangent to the frontier of the pair and to the algorithm's frontier, and at a fixed selection rate it is the horizontal distance between the two frontiers.

We next characterize the bounds on the value of the decision maker's prediction, for

[10] The second and third options echo work on algorithmic triage and learning to defer (Madras, Pitassi and Zemel, 2018; Raghu et al., 2019; Mozannar and Sontag, 2020), which elicits the decision maker's input on the cases routed to them.

which three objects are needed. The *value of perfect prediction* $V_\lambda(Y^*)$ is the value of a signal that reveals the outcome (Section 4.2), the largest λ-weighted score in the perfect-prediction set Π and the most any signal can be worth; for $\lambda \geq 1$ it equals $1-\mu$. Prediction m has an *excess reversal* if

$$p_{11}(m)\,p_{00}(m) - p_{10}(m)\,p_{01}(m) \;>\; P(M \neq m, Y^* = 1)\,P(M \neq m, Y^* = 0).$$

The left side equals $p(m)^2\,\pi(m)(1-\pi(m))\,(\mu_1(m)-\mu_0(m))$, where $\pi(m) := P(C=1 \mid M=m)$ is the selection rate at m, so an excess reversal requires a reversal within prediction m, $\mu_1(m) > \mu_0(m)$. The third object is a revealing threshold, a constant threshold at which the pair (M,H) comes arbitrarily close to revealing the outcome under some model consistent with the decision maker's choices.

Definition 4. A constant threshold t is *revealing* if, for every $\varepsilon > 0$, some threshold model consistent with the decision maker's choices with $T = t$ almost surely has $E_Q[\min\{R, 1-R\}] \leq \varepsilon$, where $R := Q(Y^* = 1 \mid M, H)$.

Lemma 3. *The revealing thresholds are the interval* $[\underline{T}, \overline{T}] \cap (\mu_1, \mu_0)$*, where*

$$\underline{T} := \max\Big\{\mu_1,\ \max_m \frac{p_{11}(m)}{p_{11}(m) + P(C=1, M \neq m, Y^*=0)}\Big\},$$
$$\overline{T} := \min\Big\{\mu_0,\ \min_m \frac{P(C=0, M \neq m, Y^*=1)}{P(C=0, M \neq m, Y^*=1) + p_{00}(m)}\Big\}.$$

There are no revealing thresholds when $\underline{T} > \overline{T}$.

Whenever we impose a restriction on their threshold, we assume that some consistent threshold model satisfies it (Corollary 2.1). We give sufficient conditions under which the data decide the upper bound without solving the programs: conditions under which it is the value of perfect prediction, and conditions under which it is informative.

Proposition 5.1. *Fix a weight* $\lambda > 1$*. Under the threshold model, the upper bound on the value of the algorithm's prediction and the decision maker's prediction is the value of perfect prediction,* $\overline{V}_\lambda(P) = 1-\mu$*, and the lower bound is the value of the algorithm's prediction together with their choice,* $\underline{V}_\lambda(P) = V_\lambda((M,C);P)$*.*

i. Monotone choice behavior. *The upper bound* $\overline{V}^{\mathrm{MC}}_\lambda(P)$ *is* $1-\mu$ *if there is no reversal across the algorithm's predictions or if some constant threshold is revealing. It is strictly below* $1-\mu$ *if some prediction has an excess reversal.*

ii. Fixed threshold. *The upper bound* $\overline{V}^{\mathrm{FT}}_\lambda(P)$ *is* $1-\mu$ *if and only if some constant threshold is revealing; otherwise it is strictly below* $1-\mu$*.*

In both cases, the lower bound remains $V_\lambda((M, C); P)$.

Consider first the lower bound. Their choice is a garbling of their prediction, so under every consistent model the pair (M, H) supports every decision rule that the pair (M, C) supports, and the value of their prediction is never below the value of their choice. The data guarantee nothing more.

Consider next the upper bound. Under the threshold model alone, it is the value of perfect prediction for every distribution of the data, for the reason given in Section 5.1: calibration holds pooled across the algorithm's predictions, so a value of their prediction may carry risky individuals at one prediction of the algorithm and safe individuals at another, and a designer who sees both separates them. Nothing in their choices rules this out.

Under a fixed threshold, the revealing thresholds decide the upper bound exactly. Every individual they select has a prediction at or below the threshold, and every individual they do not select has a prediction above it. For the pair to reveal the outcome, the risky individuals they select at prediction m must share values below the threshold with safe individuals they select at other predictions, whose mass is $P(C = 1, M \neq m, Y^* = 0)$; pooling the two groups at one value gives that value its lowest possible outcome rate, $p_{11}(m)/(p_{11}(m) + P(C = 1, M \neq m, Y^* = 0))$, and the threshold must be at least this. The safe individuals they do not select at prediction m must likewise share values above the threshold with risky individuals they do not select at other predictions, which caps the threshold at the second term of $\overline{T}$, and the averages μ_1 and μ_0 of their prediction among the selected and the unselected bound the threshold on the same two sides. A threshold outside $[\underline{T}, \overline{T}]$ leaves the individuals at some prediction without enough mass at other predictions to cover them, and Lemma 3 shows that every threshold inside it is revealing.

Monotone choice behavior allows more, because the threshold may vary with their prediction: the pair reveals the outcome under some consistent model whenever some constant threshold is revealing, and also whenever there is no reversal across the algorithm's predictions, in which case their prediction alone can nearly reveal it (Proposition 4.2). What it cannot absorb is an excess reversal. The risky individuals they select at prediction m more often than the safe ones must be covered by safe individuals they select at other predictions, who are selected at the same rate because one nonincreasing choice probability function serves every prediction, and the other predictions have only their own mass to give. The upper bound then falls strictly below the value of perfect prediction.

5.3 The Value of Judges' Input in New York

Section 4 showed that the algorithm outperforms most judges' realized choices. A judge can still add value to a designer who holds the algorithm's prediction, as long as the judge has

information that the prediction misses. We test whether the judges have such information and then bound the value of eliciting it.

5.3.1 Testing for Private Information

By Lemma 2(i), a judge has private information at a decile of predicted risk if and only if their choices and misconduct are dependent within the decile. For each judge and decile, we compare the judge's release-misconduct rate with their release rate times the decile's outcome rate, which we estimate from the other judges' decisions in the decile; the two are equal under independence. The test for a judge uses the largest studentized difference across their deciles, with standard errors from a Bayesian bootstrap with 500 draws, and because we test 371 judges we control the false discovery rate (Benjamini and Hochberg, 1995) and also report the stricter Bonferroni-Holm adjustment (Holm, 1979). Appendix H describes the tests.

Many judges that the algorithm outperforms still have private information. Table 3 reports the results by how the judge's prediction compares with the algorithm under monotone choice behavior in Table 2: 2 judges beat the algorithm, 99 cannot be ranked, the algorithm beats 122, and 139 are inconsistent with monotone choice behavior because their choices are worse than random release. Without adjusting for multiple testing, 35 percent of judges have evidence of private information, and 25 percent after the false discovery rate adjustment. The share is similar across the groups: 24 percent among the judges the algorithm beats and 24 percent among those whose choices are worse than random, against 28 percent among the judges the bounds cannot rank, and 18, 15, and 17 percent under the Bonferroni-Holm adjustment. The common comparison of Section 4.3.1 places 95 percent of judges inside the algorithm's frontier, and the bounds of Section 4.3.2 rank the algorithm above the predictions of 122 of 362 judges; for a quarter of these judges, the algorithm's prediction nevertheless does not capture everything the judge knows, and a designer who holds the prediction could do strictly better at some weight on misconduct by also eliciting the judge's choice (Lemma 2(ii)).

Panel B reports a one-sided version of the same test: whether the defendants a judge releases in a decile are riskier than those they detain, $\mu_1(m) > \mu_0(m)$, a reversal within the decile. A reversal within a decile implies a reversal across the algorithm's predictions, under which the upper bound on the judge's own information frontier lies strictly inside the perfect-prediction set under monotone choice behavior (Proposition 4.2(i)), and it is the ingredient of an excess reversal (Proposition 5.1(i)). 16% of judges reverse in some decile at the 5 percent level and 8% after the false discovery rate adjustment.

A judge's input can improve on the algorithm only in the deciles where the judge has

Table 3: Tests for Private Information and for Reversals Within Deciles, by Comparison with the Algorithm

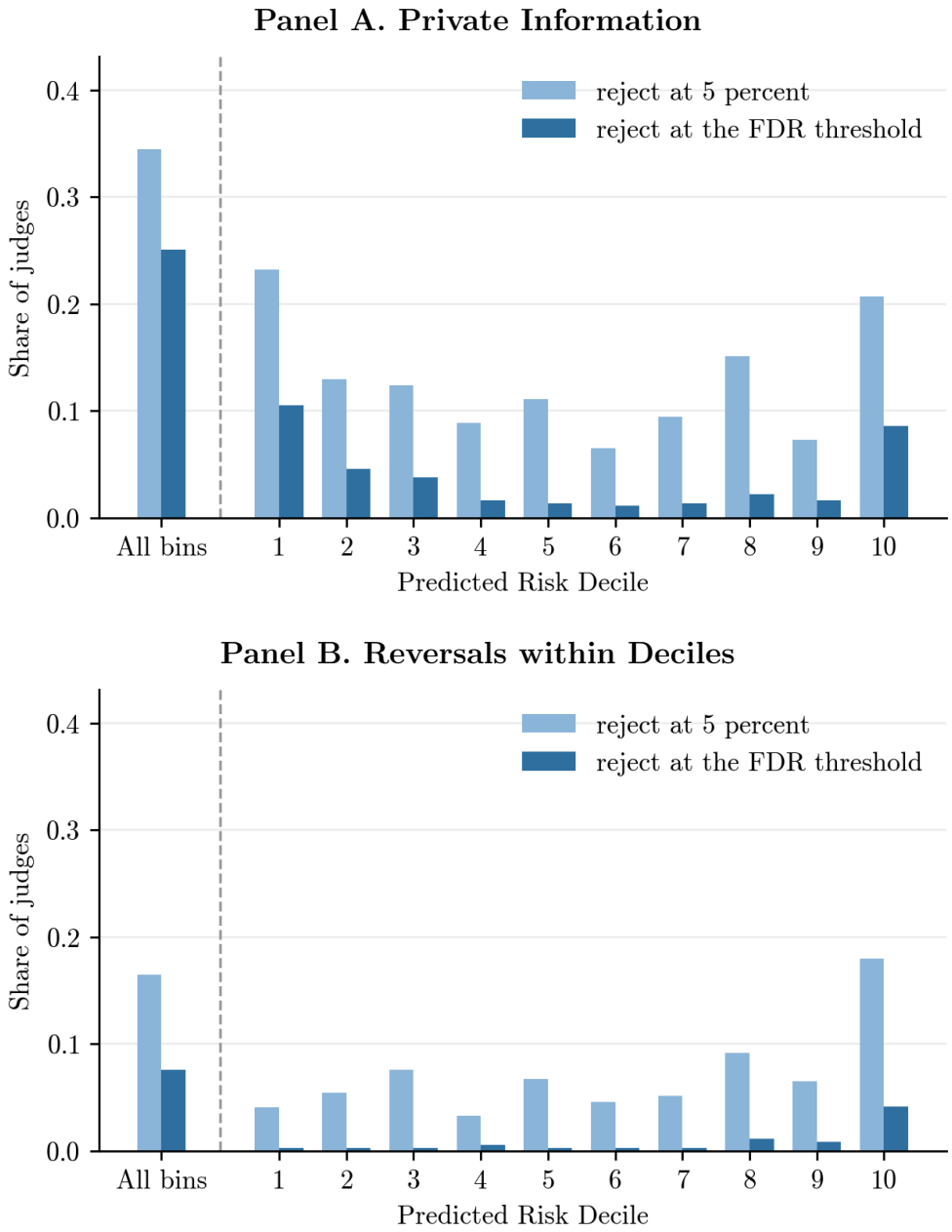
	(1)	(2)	(3)	(4)
		Share of judges whose test rejects		
			Adjusted	
Comparison with the algorithm	Judges	Unadjusted	False discovery rate	Bonferroni-Holm
Panel A: No private information				
Human $\succsim$ Algorithm	2	100	100	100
Human $\parallel$ Algorithm	99	37	28	17
Algorithm $\succsim$ Human	122	32	24	18
Inconsistent with monotone choice	139	35	24	15
All judges	362	35	25	17
Panel B: Reversals within deciles				
Human $\succsim$ Algorithm	2	0	0	0
Human $\parallel$ Algorithm	99	10	8	5
Algorithm $\succsim$ Human	122	17	6	4
Inconsistent with monotone choice	139	21	9	6
All judges	362	17	7	5

Notes. This table reports, for the 362 judges with cases in every decile of predicted risk, the percent whose test rejects at the 5 percent level, unadjusted and adjusted across all 371 tested judges, by the false discovery rate (Benjamini-Hochberg) in column (3) and by Holm's step-down procedure in column (4); column (1) counts judges. Panel A tests the hypothesis that the judge's choices are independent of misconduct within every decile, $\mu_1(m) = \mu_0(m)$ (Lemma 2(i)). Panel B tests the one-sided hypothesis that the defendants the judge releases are no riskier than those they detain within every decile, $\mu_1(m) \leq \mu_0(m)$; a rejection is a reversal within a decile. Judges are grouped by their comparison with the algorithm under monotone choice behavior, column (2) of Panel A of Table 2; the tests do not depend on the restrictions. See Section 5.3.1 for discussion.

private information. We next study where in the algorithm's predictions do judges tend to have private information. Figure 6 reports the share of judges with evidence of private information within each decile of predicted risk. The shares are highest in the extreme deciles: after the false discovery rate adjustment, 11 percent of judges in the lowest decile and 9 percent in the highest, against 1 to 2 percent in the middle deciles. More judges can add to the algorithm's prediction in the cases where its prediction is most extreme. Reversals also concentrate at the top: of the 31 judges with private information in the highest decile, 16 release riskier defendants than they detain, against 1 of 39 in the lowest, and Panel B shows the same pattern.

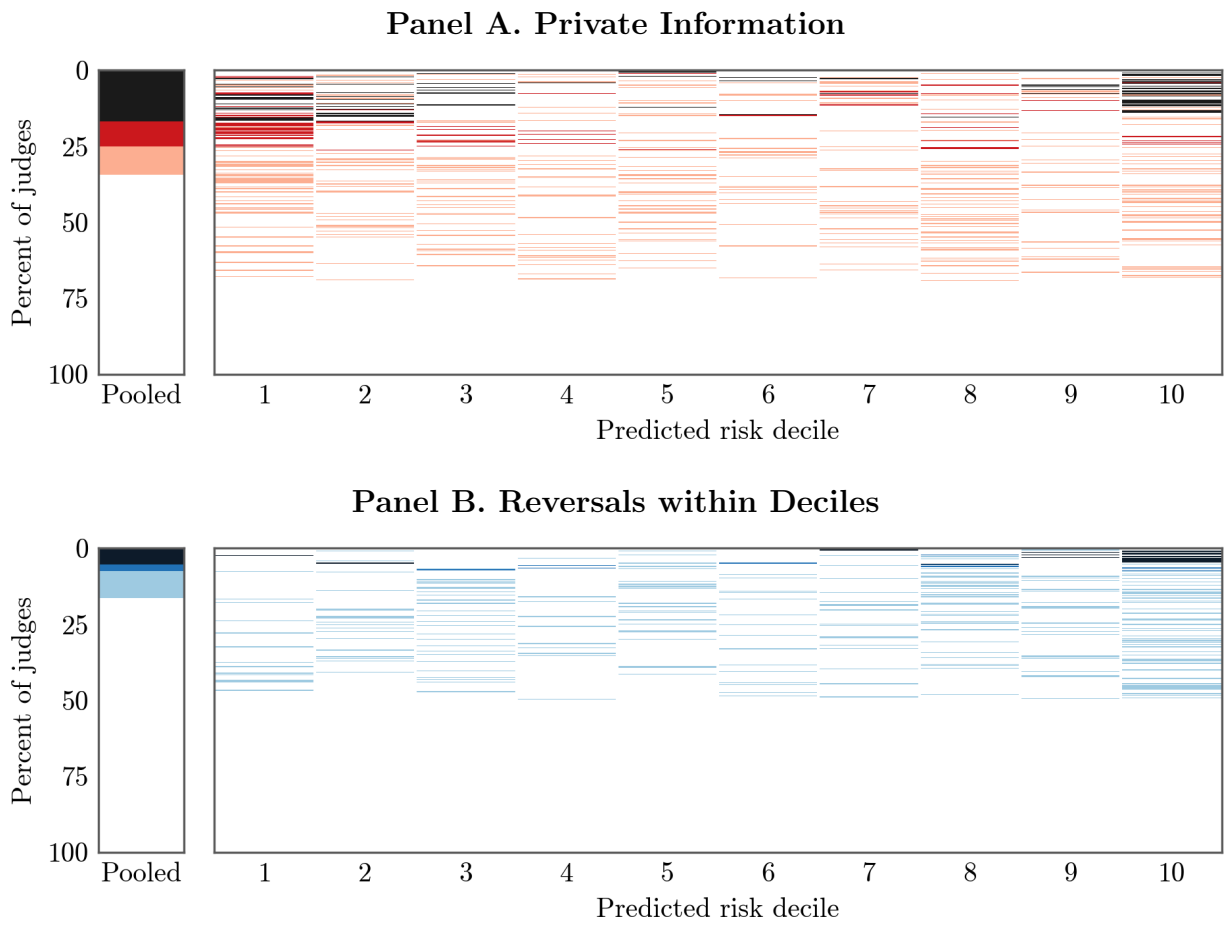
Figure 7 shows each judge and decile. Private information is not uniform: judges with it in some deciles do not have it in others, and different judges have it in different deciles, so one judge may know something the algorithm misses about low-risk defendants and another about high-risk ones. Work on AI assistance finds that assistance helps most when the algorithm's predictions are most extreme, but that decisions are better automated once the algorithm is confident enough (Agarwal et al., 2023; Agarwal, Moehring and Wolitzky, 2025). We find instead that the judges' private information is most prevalent in the cases where the algorithm is most confident.

Figure 6: Share of Judges Rejecting the Tests, by Decile of Predicted Risk



Notes. This figure reports the share of the 371 judges whose test rejects, pooled across deciles (the left bars) and within each decile of predicted risk, at the 5 percent level and after the false discovery rate adjustment. Panel A tests independence between the judge's choices and misconduct, $\mu_1(m) = \mu_0(m)$ (Lemma 2(i)); Panel B tests the one-sided hypothesis $\mu_1(m) \leq \mu_0(m)$, whose rejection is a reversal within the decile. The adjustment is across judges for the pooled test and across the 3,701judge-decile pairs for the decile tests. The two panels share the vertical scale. See Section 5.3.1 for discussion.

Figure 7: Tests for Private Information and for Reversals Within Deciles, by Judge and Decile



Color			Panel A cutoffs		Panel B cutoffs	
A	B	Test result	Pooled	Decile	Pooled	Decile
		Accept at 5%	$> 5\%$	$> 5\%$	$> 5\%$	$> 5\%$
		Reject at 5%	$\leq 5\%$	$\leq 5\%$	$\leq 5\%$	$\leq 5\%$
		Reject: False discovery rate corrected	$\leq 1.22\%$	$\leq 0.16\%$	$\leq 0.37\%$	$\leq 0.04\%$
		Reject: Bonferroni-Holm corrected	$\leq 0.01\%$	$\leq 0.001\%$	$\leq 0.004\%$	$\leq 0.0002\%$

Notes. This figure reports, for each of the 371 judges (rows) and each decile of predicted risk (columns), the result of the test for private information (Panel A) and of the one-sided test for a reversal within the decile (Panel B). In each panel the narrow strip on the left is the pooled test across deciles, one block per judge, and the grid on the right is the test within each decile. Judges are sorted by their pooled p-value, most significant first, so a colored block in the pooled strip reads as the percent of judges rejecting at that level; each panel sorts by its own test, so a row is not the same judge across panels. The key gives each color's test result and its p-value cutoff; the adjustments are across judges for the pooled test and across judge-decile pairs for the decile tests. A square is white where the judge has no cases in that decile. The standard errors come from a Bayesian bootstrap with 500 draws. See Section 5.3.1 for discussion.

5.3.2 Bounding the Value of Eliciting Judges' Input

A designer who holds the algorithm's prediction can elicit a judge's choice, made without the algorithm, or the judge's prediction. The value of the algorithm's prediction with the judge's choice is identified from the judge's strata-adjusted release rate and release-misconduct rate within each decile, as in Section 4.3.2, and the value with the judge's prediction is bounded by Proposition 5.1. Because the designer's weight on misconduct is unknown, we fix the comparison at the judge's own release rate. At that release rate, we compare the release-misconduct rate of the algorithm, of the algorithm with the judge's choice, and of the algorithm with the judge's prediction. The first two are identified; for the third we report the lowest rate the data cannot rule out, the bound $\underline{\tilde{S}}_{(M,H)}(r; P)$ of Section 5.2 under each restriction, computed with the programs of Appendix C.2.[11]

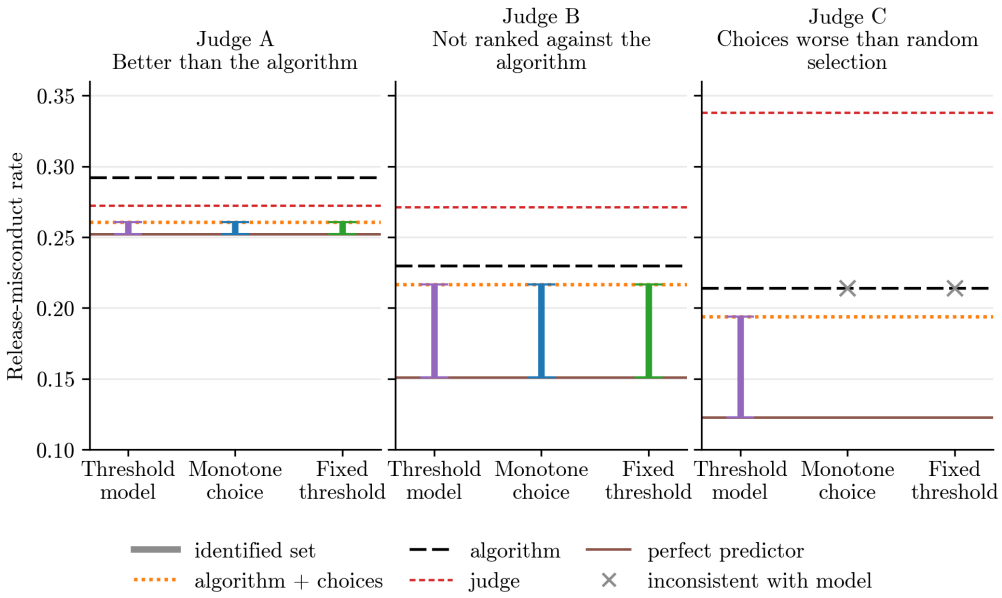
Figure 8 presents the bounds at the release rates of Judges A, B, and C from Section 4.3.2, under each restriction consistent with the judge's choices. Each bar runs from the release-misconduct rate of the algorithm with the judge's choice to the lowest rate that the algorithm with the judge's prediction could achieve. For Judge A, who beats the algorithm, the algorithm with the judge's choice comes close to the perfect predictor and the bounds are tight. For Judge B, whom the algorithm beats, eliciting the choice still lowers the algorithm's rate, and the bound runs to the perfect predictor: under monotone choice behavior Judge B's prediction alone cannot reach the perfect predictor (Figure 5), but combined with the algorithm it can. For Judge C, consistent with the threshold model alone, eliciting the choice lowers the rate a little, and the bound again runs to the perfect predictor.

Figure 9 traces the same comparison across release rates. For all three judges, the frontier of the algorithm with the judge's choice lies above the algorithm's at most release rates, and the upper bound is the frontier of the perfect predictor under the threshold model alone and nearly so under the two restrictions. The figure also shows that the judges' choices are not a garbling of the algorithm's prediction, nor the prediction of the choices: in the first case the frontier of the algorithm with the choice would coincide with the algorithm's, and in the second the algorithm with the choice would do no better than the choice alone, whereas for all three judges it does better than either at the judge's release rate.

Table 4 summarizes the comparison at each judge's own release rate for the judges consistent with each restriction: the release-misconduct rates of random release, of the algorithm, of the algorithm with the judge's choice, and of the upper bound for the algorithm with the judge's prediction, and the reduction relative to the algorithm from each. Eliciting the

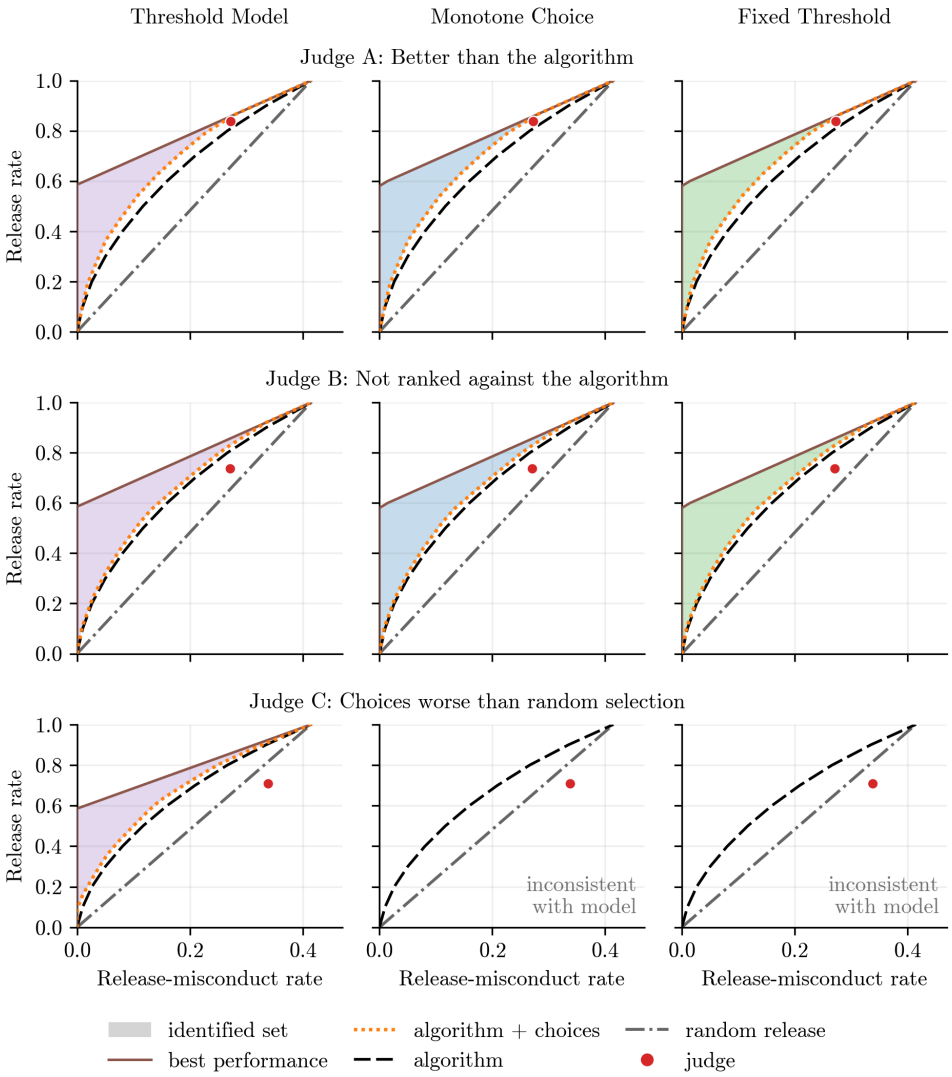
[11] Our analysis assumes that judges report their choices and their predictions as they would without the algorithm; of course, a judge who knows that their input will be combined with the algorithm may respond strategically, which would require a mechanism design approach.

Figure 8: Bounds on the Value of Three Judges' Input at Their Release Rates



Notes. This figure reports, for Judges A, B, and C of Section 4.3.2, the release-misconduct rates at each judge's own release rate of the algorithm (black dashed line), of the algorithm with the judge's choice (orange dotted line), and of the bounds on the algorithm with the judge's prediction. Each panel is one judge, and each bar is the identified set for the algorithm with the judge's prediction under the restrictions on the horizontal axis, cumulative from left to right (Proposition 5.1): its top is the algorithm with the judge's choice and its bottom the perfect predictor (solid brown line) under every restriction. The red dashed line is the judge's realized rate. Deciles are weighted equally, and the bars are point estimates. See Section 5.3.2 for discussion.

Figure 9: Bounds on the Frontier of the Algorithm with Three Judges' Input



Notes. This figure reports, for Judges A, B, and C of Section 4.3.2, the bounds on the information frontier of the algorithm with the judge's prediction at every release rate. Each row is one judge and each column one combination of restrictions, cumulative from left to right. The shaded region runs from the upper bound (solid line) to the frontier of the algorithm with the judge's choice (orange dotted line), the lower bound. Under the two restrictions, the upper bound solves the programs of Appendix C.2 at release rates in steps of 0.02 for Judges A and B, and every solve reaches the perfect predictor; Judge C's choices are inconsistent with those restrictions. Deciles are weighted equally, and the bounds are point estimates. See Section 5.3.2 for discussion.

judges' choices lowers the release-misconduct rate under every restriction: by at least 0.9 percentage points from 26.0 percent for the algorithm under the threshold model alone, and by at least 1.1 percentage points among the 223 judges consistent with the two restrictions. We cannot rule out much larger gains from eliciting the judges' prediction: in every column, the upper bound lowers the release-misconduct rate by 5.7 to 5.8 percentage points, more than a fifth of the algorithm's rate, and it is insensitive to the model restrictions. We find that no judge has an excess reversal. An excess reversal requires a prediction to hold a large share of the population, more than half when the outcome rate at the other predictions is one half, and with the algorithm's prediction coarsened to deciles no decile holds more than a tenth.

Table 4: Bounds on the Value of the Judges' Input at Their Own Release Rates

	(1) Threshold Model	(2) Monotone Choice	(3) Fixed Threshold
Judges consistent with the model	362	223	223
Panel A: Release-misconduct rate at the judge's release rate			
Random release	32.7	32.6	32.6
	[29.2, 35.9]	[29.5, 35.6]	[29.5, 35.6]
Algorithm	26.0	25.8	25.8
	[21.2, 31.2]	[21.6, 30.6]	[21.6, 30.6]
Algorithm + choices	25.1	24.7	24.7
	[20.0, 30.4]	[20.3, 29.5]	[20.3, 29.5]
Algorithm + human information, upper bound	20.3	20.0	20.0
	[11.9, 28.2]	[12.6, 27.3]	[12.6, 27.3]
Panel B: Value of elicitation			
Algorithm + choices	0.9	1.1	1.1
	[0.1, 2.2]	[0.2, 2.3]	[0.2, 2.3]
Algorithm + human information, upper bound	5.7	5.8	5.8
	[3.1, 9.3]	[3.4, 9.0]	[3.4, 9.0]

Notes. This table reports, for the 362 judges with cases in every decile of predicted risk, the release-misconduct rate at the judge's own release rate of random release, of the algorithm, of the algorithm with the judge's choice, and of the upper bound for the algorithm with the judge's prediction, as medians across the judges consistent with each restriction, in percent (Panel A), and the reduction relative to the algorithm from each, in percentage points (Panel B). The algorithm with the judge's choice is the lower bound on the value of eliciting the judge's prediction under every restriction, and the perfect predictor is its upper bound. Deciles are weighted equally. See Section 5.3.2 for discussion.

6 The Value of AI Assistance

We finally turn to the third question: how much better off would the decision maker be if they also observed the algorithm's prediction? The value of AI assistance is the gain from the algorithm's prediction at the weights the decision maker's own threshold implies, and it is

only partially identified. We bound it under the threshold model and under each restriction on their threshold. We then estimate the value of AI assistance for the judges in New York pretrial release.

6.1 Bounding the Value of AI Assistance under the Threshold Model

Consider a decision maker who observes the algorithm's prediction. The prediction expands their information from their prediction H to the pair (M, H), while their preferences remain their own: a threshold t corresponds to the weight $1/t$ on the two components of the score. The value of AI assistance is the value of the pair over the value of their prediction alone at the weights their threshold implies. We judge it by the decision maker's own objective, which is the appropriate metric if their preferences summarize socially relevant considerations.

For a weight $\lambda > 0$, the gain from adding the algorithm's prediction to the decision maker's prediction is

$$\Delta_\lambda(Q) := V_\lambda\big((M, H); Q\big) - V_\lambda(H; Q),$$

which is nonnegative by construction. Their threshold is a random variable, and so we define the *value of AI assistance* as

$$\Delta(Q) := E_Q\big[\Delta_{1/T}(Q)\big]. \tag{7}$$

Since $T \in (0, 1]$, every weight they use is at least one, and the gain is zero wherever their threshold equals one (Section 5.2). The definition evaluates the gain at their weights rather than at a weight chosen by the designer. Under a fixed threshold $T = t$, the value of AI assistance is simply $\Delta(Q) = \Delta_{1/t}(Q)$. The value of AI assistance is the gain when the decision maker makes full use of the algorithm's prediction: their weights remain their own, and the pair is used as well as it can be.[12]

Because every weight they use is at least one, the value of perfect prediction is $1 - \mu$ at every one of them, and the value of AI assistance lies in $[0, 1-\mu]$ under every consistent model. The two primitives of Proposition 2.1 determine the joint distribution of (M, Y^*, H, C), but they pin the distribution of their threshold only where it binds: they fix the choice probability function ρ_Q of Section 2.3, not where the threshold lies on either side of their prediction, and $\Delta(Q)$ depends on where it lies. Two threshold models with the same distribution of (M, Y^*, H, C) can therefore differ in the value of AI assistance. We define the upper bound

[12] How much of this gain they realize in practice, for example whether they attend to the prediction, is a separate question, which deployments and experiments can address and observational data cannot. Theoretical work on AI assistance shares this structure: the decision maker responds to a provided signal, correctly at the benchmark and with measured deviations otherwise (Dreyfuss and Hoong, 2026; Agarwal, Moehring and Wolitzky, 2025).

as the supremum over the identified set $\mathcal{Q}(P)$, the most the data cannot rule out, and the lower bound as the infimum, what the data guarantee; substituting $\mathcal{Q}^{\mathrm{MC}}(P)$ or $\mathcal{Q}^{\mathrm{FT}}(P)$ gives the bounds under monotone choice behavior and under a fixed threshold.

The bounds are characterized by optimization: each is the optimal value of a program in the primitives of Section 2.3 together with the distribution of their threshold at each value of their prediction, which the data pin only on the side their choice determines (Appendix C.2). When we report the value at the decision maker's own selection rate, as in Section 5.2, the weight is replaced by a constraint on the selection rate: the value is the reduction in the selected-outcome rate from the frontier of their prediction alone at rate r, $\tilde{S}_H(r;Q)$, to the frontier of the pair, $\tilde{S}_{(M,H)}(r;Q)$, at their own selection rate, and its bounds are the optimal values of the same programs. Appendix Figure A3 illustrates both measures with the roles of the two predictions exchanged: at a fixed selection rate the value is the horizontal distance between the frontier of their prediction and the frontier of the pair, and at a threshold t it is the vertical distance between the lines of slope $1/t$ tangent to the two frontiers.

The next result gives the lower bound and the condition, in terms of the revealing thresholds of Lemma 3, under which the upper bound is the value of perfect prediction.

Proposition 6.1. *Under the threshold model, the lower bound on the value of AI assistance is zero, and the upper bound is the value of perfect prediction, $1-\mu$, if and only if μ_1 is a revealing threshold.*

i. Monotone choice behavior. *The bounds are the same: the lower bound is zero, and the upper bound is $1-\mu$ if and only if μ_1 is a revealing threshold.*

ii. Fixed threshold. *The lower bound is*

$$\min_{t\in[\mu_1,\mu_0]}\Big(V_{1/t}\big((M,C);P\big)-V_{1/t}(C;P)\Big),$$

which is positive if and only if $\max_m \mu_1(m) > \min_m \mu_0(m)$. The upper bound is $1-\mu$ if and only if μ_1 is a revealing threshold.

Consider first the lower bound, which is what the data guarantee. Under the threshold model alone, they guarantee nothing. Their threshold is pinned only where it binds, so a consistent model may place it at one wherever they select, where the algorithm's prediction is worth nothing to them, and far below every outcome probability wherever they do not, where neither their prediction nor the pair is worth anything; under such a model the value of AI assistance is zero. Monotone choice behavior restricts only the probability that their threshold is at least their prediction, which this model respects, so it adds nothing here,

in contrast with Propositions 4.2 and 5.1. Under a fixed threshold, the guarantee is exact. Their threshold is a constant $t \in [\mu_1, \mu_0)$ (Corollary 2.1), and at the weight $1/t$ the value of their prediction equals the value of their choice, so the value of AI assistance is at least the gain from adding the algorithm's prediction to their choice at that weight, with equality when their prediction is a function of their choice. The lower bound is the smallest such gain over the thresholds they may be using, and it is positive under a condition weaker than a reversal within a prediction: the individuals they select at some prediction need only be riskier than those they do not select at some other.

Consider next the upper bound, which is what the data cannot rule out. The condition is the same under every restriction, because the weight is set by the model rather than by the designer: restrictions on their threshold constrain where it lies but not what a value of their prediction means at each prediction of the algorithm, and the data decide the upper bound through the latter. At a fixed threshold t, their prediction is worth $P(C = 1)(1 - \mu_1/t)$ to them, which vanishes as t falls to μ_1. When μ_1 is a revealing threshold, the data cannot rule out models with a fixed threshold just above μ_1 under which their prediction is worth almost nothing to them while the pair nearly reveals the outcome; along such a sequence the value of AI assistance approaches the value of perfect prediction, which no consistent model attains (Appendix A.3). When μ_1 is not a revealing threshold, no consistent model comes close: for the pair to reveal the outcome their threshold must sit away from μ_1, where their prediction alone is already worth something to them, and the upper bound falls strictly below the value of perfect prediction.

6.2 The Value of AI Assistance in New York

A judge who observes the algorithm's prediction keeps the decision and the judge's own weight on misconduct (Section 6.1). Because the judge's threshold is not identified, we report the value in two ways. At the judge's own release rate, the value is the reduction in the release-misconduct rate from the judge's prediction alone to the judge's prediction together with the algorithm's. At the judge's own thresholds, it is the expected gain $\Delta(Q)$, which we report as the equivalent reduction in the release-misconduct rate at an unchanged release rate, multiplying the gain at a threshold t by t. Under a fixed threshold, the bounds use the least and the most favorable t in the interval of Corollary 2.1. Both measures compare two partially identified signals, and we report their bounds under each restriction consistent with the judge's choices, computed with the programs of Appendix C.3. Each bound is a change in the judge's release-misconduct rate, with the judge's own decisions at zero.

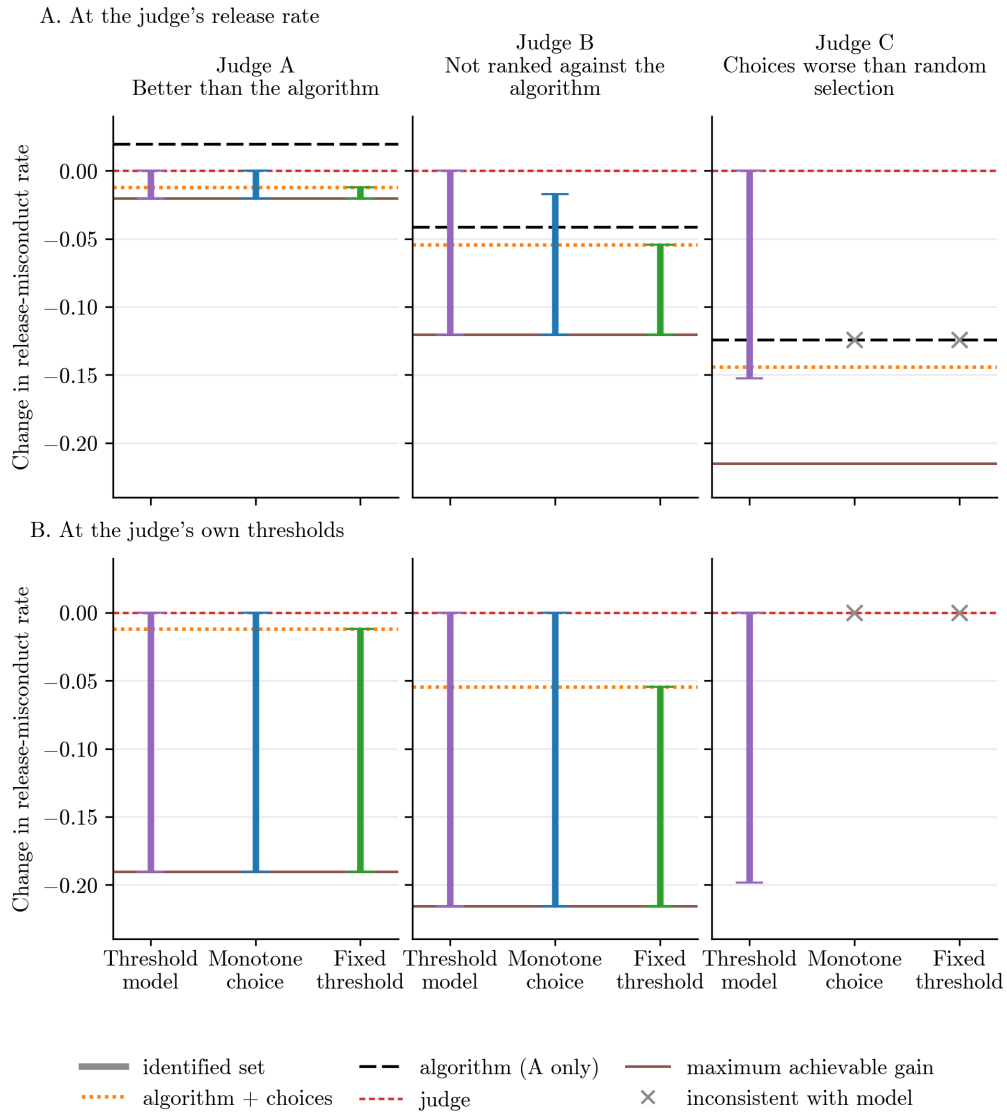
Figure 10 presents the bounds for Judges A, B, and C of Section 4.3.2, at the judge's release rate in Panel A and at the judge's own thresholds in Panel B, under each restriction

consistent with the judge's choices. Each bar runs from the smallest to the largest gain the data allow. For Judge A, who beats the algorithm, the gain at the judge's release rate is at most 2.0 percentage points, the gap between the judge's realized rate and the perfect predictor, and it can be zero under the threshold model alone and under monotone choice behavior, because Judge A's prediction could be as good as the perfect predictor (Section 4.3.2); under a fixed threshold the smallest gain is the gain from adding the algorithm's prediction to the judge's choice. For Judge B, whom the algorithm beats, the lower bound at the judge's release rate is positive already under monotone choice behavior, because Judge B's prediction cannot reach the perfect predictor (Section 4.3.2), and the upper bound is 12.0 percentage points. For Judge C, consistent with the threshold model alone, the upper bound at the judge's release rate falls short of the perfect predictor, and at the judge's own thresholds it lies below the value of perfect prediction because μ_1 is not a revealing threshold for this judge (Proposition 6.1). At the judge's own thresholds, the judge may also change the release rate; the lower bound is zero except under a fixed threshold, where it is again the gain from adding the algorithm's prediction to the judge's choice, and the upper bounds are larger, 19 to 22 percentage points.

Figure 11 traces the value for Judges A and B under monotone choice behavior across every release-misconduct rate that the judge's prediction alone could achieve at the judge's release rate, which the data do not identify. The gain is zero only if the judge's prediction alone does as well as the pair at that rate. Monotone choice behavior allows this for Judge A: if Judge A's prediction achieves a release-misconduct rate of 25.2 percent, it matches the perfect predictor and the algorithm's prediction adds nothing, and if it is no better than the judge's choices, the algorithm's prediction lowers the rate by at least 4.4 percent. It does not allow it for Judge B, whose prediction cannot reach the perfect predictor (Figure 5): the algorithm's prediction lowers the rate by at least 8.7 percent across the whole range and by at least 20.1 percent when the judge's prediction is no better than the judge's choices.

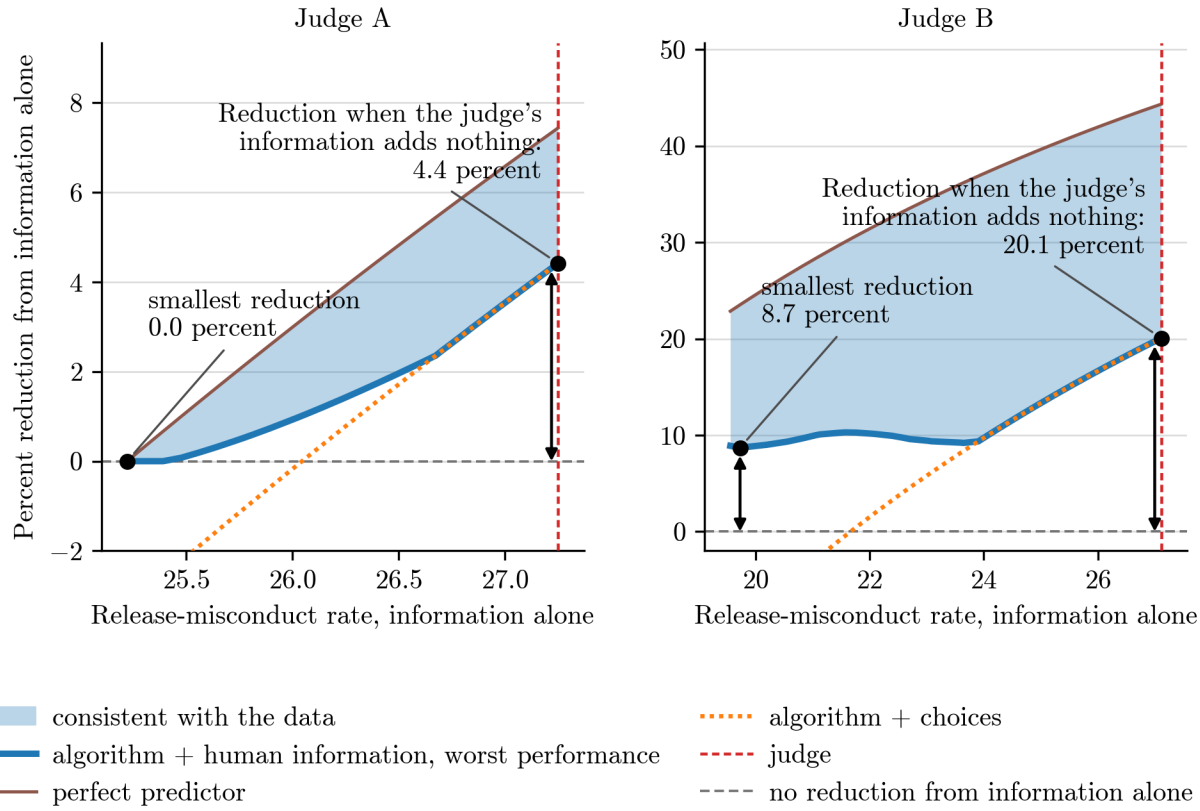
Table 5 summarizes the bounds for the 223 judges consistent with a fixed threshold, as means across judges with 90 percent credible intervals from 500 Bayesian bootstrap draws. Under the threshold model alone, the data do not rule out that AI assistance has no value, because each judge's prediction could be a perfect predictor without the algorithm. The upper bound is large, 10.7 percentage points at the judge's release rate, 36.0 percent of the judge's rate, and 22.9 percentage points at the judge's own thresholds, 75.0 percent of the judge's rate, and it is the same under every restriction because μ_1 is a revealing threshold for every one of these judges: with the algorithm's prediction coarsened to ten equal-weight deciles, the decile terms of $\underline{T}$ and $\overline{T}$ do not bind, and $\mu_1 < \mu_0$ holds for every judge consistent with a fixed threshold (Corollary 2.1). Monotone choice behavior makes the lower bound

Figure 10: Bounds on the Value of AI Assistance for Three Judges at Their Release Rates and at Their Own Thresholds



Notes. This figure reports, for Judges A, B, and C of Section 4.3.2, the bounds on the value of giving the judge the algorithm's prediction, as the change in the judge's release-misconduct rate, under the restrictions on the horizontal axis, cumulative from left to right; a cross marks a restriction inconsistent with the judge's choices. Panel A holds the judge's release rate fixed. Panel B evaluates the gain at the judge's own thresholds, which under a fixed threshold lie between the misconduct rates of the judge's released and detained defendants, and reports it as the equivalent change at an unchanged release rate. Each bar is the identified set, and lower is better. The horizontal lines mark the judge's own decisions (red dashed line, at zero), the algorithm (black dashed line, Panel A only), the algorithm with the judge's choice (orange dotted line), and the perfect predictor (solid brown line). Judge C's bound in Panel B is the best value found numerically. Deciles are weighted equally, and the bars are point estimates. See Section 6.2 for discussion.

Figure 11: The Value of AI Assistance Across the Possible Performance of Two Judges' Predictions



Notes. This figure reports, for Judges A and B under monotone choice behavior, the value of giving the judge the algorithm's prediction at the judge's release rate, as a percent of the release-misconduct rate of the judge's prediction alone (vertical axis), against that rate (horizontal axis), which the data do not identify. The shaded region collects the pairs consistent with the judge's choices, from the lowest value the pair could achieve (solid blue line) to the perfect predictor (solid brown line). The orange dotted line is the algorithm with the judge's choice, and the red dashed vertical line marks the judge's realized rate. The left dot marks the smallest reduction, and the right dot the reduction when the judge's prediction adds nothing to the judge's choice, the lower bound under a fixed threshold in Panel A of Figure 10. Deciles are weighted equally, and the bounds are point estimates. See Section 6.2 for discussion.

positive at the judge's release rate, 3.4 percentage points on average and positive for 182 judges, and leaves it at zero at the judge's own thresholds. A fixed threshold gives the largest lower bounds, 5.9 percentage points at the judge's release rate and 7.2 at the judge's own thresholds, positive for all 223 judges. A fixed threshold is also the restriction under which the common comparison measures predictive skill and under which the algorithm beats 95 percent of judges at their release rates (Section 4.3.1); the same restriction gives the largest guaranteed value from giving judges the algorithm's prediction, so automation is not the only response to judges who perform worse than the algorithm.

Table 5: Bounds on the Value of AI Assistance for the Judges at Their Release Rates and at Their Own Thresholds

	(1) Threshold Model	(2) Monotone Choice	(3) Fixed Threshold
Panel A: at the judge's release rate			
Percentage points			
Lower bound	0.0	3.4	5.9
	[0.0, 0.0]		[5.6, 6.3]
Upper bound	10.7	10.7	10.7
	[9.9, 11.0]	[9.9, 11.0]	[9.9, 11.0]
Percent of the judge's release-misconduct rate			
Lower bound	0.0	11.2	19.5
	[0.0, 0.0]		[18.7, 21.2]
Upper bound	36.0	36.0	36.0
	[33.7, 38.1]	[33.7, 38.1]	[33.7, 38.1]
Panel B: at the judge's own thresholds			
Percentage points			
Lower bound	0.0	0.0	7.2
	[0.0, 0.0]	[0.0, 0.0]	[6.7, 7.6]
Upper bound	22.9	22.9	22.9
	[22.4, 22.8]	[22.4, 22.8]	[22.4, 22.8]
Percent of the judge's release-misconduct rate			
Lower bound	0.0	0.0	23.5
	[0.0, 0.0]	[0.0, 0.0]	[22.1, 25.2]
Upper bound	75.0	75.0	75.0
	[74.3, 76.1]	[74.3, 76.1]	[74.3, 76.1]
Judges with a positive lower bound			
At the judge's release rate	0	182	223
At the judge's own thresholds	0	0	223

Notes. This table reports, for the 223judges consistent with a fixed threshold, the bounds on the value of giving the judge the algorithm's prediction under each restriction (columns), as means across judges of the reduction in the judge's release-misconduct rate, in percentage points and as a percent of the judge's rate, and the number of judges with a positive lower bound. Panel A holds the judge's release rate fixed. Panel B evaluates the gain at the judge's own thresholds, which under a fixed threshold lie between the misconduct rates of the judge's released and detained defendants, and reports it as the equivalent reduction at an unchanged release rate. Brackets give 90 percent credible intervals from 500 Bayesian bootstrap draws; the lower bound under monotone choice behavior in Panel A is a point estimate only. Deciles are weighted equally. See Section 6.2 for discussion.

7 Conclusion

Algorithms now outperform many human decision makers on aggregate performance metrics, and these comparisons often guide which decisions to automate. We develop a behavioral model of binary classification in which each choice applies a threshold to the decision maker's prediction of the outcome. The model attributes every factor in their choices that is correlated with the outcome to their prediction and everything else to their threshold. On its own the model has no testable implications, and the two restrictions on their threshold that we study, monotone choice behavior and a fixed threshold, are testable through a single comparison: the individuals the decision maker selects must be less risky on average than those they do not.

Using the threshold model, we first ask whether the decision maker's predictive skill compares favorably with the algorithm's. Comparing their choices with the algorithm's frontier measures their predictive skill only under a fixed threshold. Under monotone choice behavior, the data instead bound their information frontier, and the algorithm can be shown to outperform their prediction only when there is a reversal across its predictions. In New York pretrial release, replacing a fixed threshold with monotone choice behavior cuts the share of judges whom the algorithm is estimated to beat by approximately half. We next ask whether a designer who observes the algorithm's prediction gains by also eliciting the decision maker's choice or their prediction. The decision maker has private information if and only if their choice is correlated with the outcome within some prediction of the algorithm, and we show that the value of eliciting their prediction is bounded. For about a quarter of judges in New York, we reject the null hypothesis of no private information, so a designer who holds the algorithm's prediction would gain by also eliciting their choices. We finally ask how much better off the decision maker would be if they also observed the algorithm's prediction, at the weights their own threshold implies. We estimate that providing judges with the algorithm's prediction could lower their release-misconduct rates at their release rates by at least 5.9 percentage points on average.

The common comparison of humans and algorithms is meant to inform the design of high-stakes decisions, yet it answers none of the questions we ask. It does not measure the decision maker's predictive skill in general, because their choices combine their prediction with a threshold that may vary across individuals. Nor does it reveal whether the decision maker adds anything to the algorithm: a decision maker whose choices lie inside the algorithm's frontier may still hold a prediction that is informative given the algorithm's. In New York, the algorithm outperforms many judges who nevertheless have private information, and combining the two improves on the algorithm alone. More broadly, the question for a designer

is how best to combine the decision maker and the algorithm where each has a comparative advantage, and the data behind the common comparison can already bound the answer. Under the fixed threshold that existing structural models impose (Chan, Gentzkow and Yu, 2022; Arnold, Dobbie and Hull, 2022; Ribers and Ullrich, 2023; Rambachan, 2024), the restriction under which the common comparison measures predictive skill, the same data also bound the value of eliciting the decision maker's input and the value of AI assistance. Automation is therefore not the only response to a decision maker who performs worse than the algorithm: the assumption that favors it also bounds the value of the alternatives.

Our analysis has several limitations that motivate further work. First, the data record the decision maker's choices rather than their predictions. A choice reveals only whether their prediction lies above or below their threshold, and much of what our bounds leave open comes from this. Where decision makers' predictions can be elicited (as in Agarwal et al., 2023; Agarwal, Moehring and Wolitzky, 2025), they identify the information frontier and the value of their input directly. Second, a large share of judges make choices that are inconsistent with monotone choice behavior, and the threshold model alone rationalizes them. How these judges decide is an open question. Calibration also holds pooled across the algorithm's predictions, so our test for private information is one-directional: evidence of private information is robust, but its absence does not show that the decision maker's prediction is uninformative. A restriction that the outcome probability within each prediction of the algorithm is nondecreasing in the decision maker's prediction would close this gap, and we leave its testable content and its consequences for our bounds to future work. Our bounds also rely on the algorithm's prediction, which is informative because it is correlated with the decision maker's prediction. Variables that shift their threshold but not their prediction would identify more. Finally, the value of AI assistance is defined under full use of the algorithm's prediction, so our bounds are bounds on that benchmark. Prior work suggests that decision makers do not use predictions fully (Agarwal et al., 2023).

While we applied it to analyze pretrial release decisions in New York, we provide a framework for understanding how to combine humans and algorithms based on the data that the common comparison already uses. Our framework can be applied across empirical settings that compare humans and algorithms, such as hiring (e.g., Hoffman, Kahn and Li, 2017), medical testing (e.g., Mullainathan and Obermeyer, 2021), loan approval (e.g., Dobbie et al., 2021), child welfare screening (e.g., Baron et al., 2026), and many others (Kleinberg et al., 2015; Mullainathan, 2025). Within binary classification, how decision makers respond once their input is used to make decisions, and the value they add by formulating the problem, lie outside our framework. Algorithms have also moved beyond binary classification, and many now take or produce natural language. These models are ranked on benchmarks that

aggregate performance across component problems, much like the classification metrics we study. Where preferences over outputs vary across the problems in a benchmark, rankings by its aggregate score may face the same issues we find for classification.

References

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh. 2016. "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care." *American Economic Review*, 106(12): 3730–64.

Agarwal, Nikhil, Alex Moehring, and Alexander Wolitzky. 2025. "Designing Human-AI Collaboration: A Sufficient-Statistic Approach." National Bureau of Economic Research Working Paper 33949.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." National Bureau of Economic Research Working Paper 31422. Revise and resubmit, *Econometrica*.

Alur, Rohan, Loren Laine, Darrick K. Li, Manish Raghavan, Devavrat Shah, and Dennis Shung. 2023. "Auditing for Human Expertise."

Alur, Rohan, Manish Raghavan, and Devavrat Shah. 2024. "Human Expertise in Algorithmic Prediction."

Angelova, Victoria, Will S Dobbie, and Crystal Yang. 2023. "Algorithmic Recommendations and Human Discretion." National Bureau of Economic Research Working Paper 31747.

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

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

Baron, E. Jason, Arkadev Ghosh, and Richard Lombardo. 2026. "Predictive Risk Scores in the Public Sector: Experimental Evidence from Child-Protection Investigations." National Bureau of Economic Research Working Paper 35540.

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

Baron, E. Jason, Will Dobbie, Richard Lombardo, Ashesh Rambachan, and Joseph Ryan. 2026. "Human Decisions and Machine Predictions in Multistage Systems." Working paper.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300.

Blackwell, David. 1953. "Equivalent Comparisons of Experiments." *The Annals of Mathematical Statistics*, 24(2): 265–272.

Bonhomme, Stéphane, and Martin Weidner. 2021. "Minimizing Sensitivity to Model Misspecification."

Canay, Ivan A., Magne Mogstad, and Jack Mountjoy. 2024. "On the Use of Outcome Tests for Detecting Bias in Decision Making." *Review of Economic Studies*, 91(4): 2135–2167.

Caplin, Andrew, and Daniel Martin. 2015. "A Testable Theory of Imperfect Perception." *The Economic Journal*, 125(582): 184–202.

Caplin, Andrew, and Mark Dean. 2015. "Revealed Preference, Rational Inattention, and Costly Information Acquisition." *American Economic Review*, 105(7): 2183–2203.

Chamberlain, Gary, and Guido W. Imbens. 2003. "Nonparametric Applications of Bayesian Inference." *Journal of Business & Economic Statistics*, 21(1): 12–18.

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

Cowgill, Bo. 2018. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening." Columbia Business School Working Paper.

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

Dreyfuss, Bnaya, and Ruru Hoong. 2026. "Calibrated Coarsening: Designing Information for AI-Assisted Decisions." Working paper.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2021. "Predictably Unequal? the Effects of Machine Learning on Credit Markets." *The Journal of Finance*, 77(1): 5–47.

Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li. 2017. "Discretion in Hiring." *The Quarterly Journal of Economics*, 133(2): 765–800.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, 6(2): 65–70.

Hull, Peter. 2021. "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making." National Bureau of Economic Research Working Paper 28503.

Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(2): 167–189.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: a Flaw in Human Judgment.* . First ed., New York:Little, Brown Spark.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics*, 133(1): 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review*, 105(5): 491–95.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." 275–284.

Li, Danielle, Lindsey R. Raymond, and Peter Bergman. 2020. "Hiring as Exploration." National Bureau of Economic Research Working Paper 27736.

Liu, Xiaoxuan, Livia Faes, Aditya U. Kale, Siegfried K. Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R. Ledsam, Martin K. Schmid, Konstantinos Balaskas, Eric J. Topol, Lucas M. Bachmann, Pearse A. Keane, and Alastair K. Denniston. 2019. "A Comparison of Deep Learning Performance Against Health-Care

Professionals in Detecting Diseases From Medical Imaging: a Systematic Review and Meta-Analysis." *The Lancet Digital Health*, 1(6): e271–e297.

Madras, David, Toniann Pitassi, and Richard Zemel. 2018. "Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer." Vol. 31.

McLaughlin, Bryce, and Jann Spiess. 2023. "Algorithmic Assistance with Recommendation-Dependent Preferences." 991. Full paper: arXiv:2208.07626.

Mozannar, Hussein, and David Sontag. 2020. "Consistent Estimators for Learning to Defer to an Expert." Vol. 119 of *Proceedings of Machine Learning Research*, 7076–7087.

Mullainathan, Sendhil. 2025. "Economics in the Age of Algorithms." *AEA Papers and Proceedings*, 115: 1–23.

Mullainathan, Sendhil, and Ziad Obermeyer. 2021. "Diagnosing Physician Error: a Machine Learning Approach to Low-Value Health Care." *The Quarterly Journal of Economics*, 137(2): 679–727.

Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. "The Algorithmic Automation Problem: Prediction, Triage, and Human Effort." arXiv:1903.12220.

Rambachan, Ashesh. 2024. "Identifying Prediction Mistakes in Observational Data." *The Quarterly Journal of Economics*, 139(3): 1665–1711.

Ribers, Michael Allan, and Hannes Ullrich. 2023. "Machine Predictions and Human Decisions with Variation in Payoff and Skills: the Case of Antibiotic Prescribing." Berlin School of Economics Working Paper 0027.

Rodriguez, Krystal, and Michael Rempel. 2023. "Explaining New York's 2023 Bail Reform Amendments." Data Collaborative for Justice at John Jay College.

Rubin, Donald B. 1981. "The Bayesian Bootstrap." *Annals of Statistics*, 9(1): 130–134.

Stevenson, Megan T., and Jennifer L. Doleac. 2024. "Algorithmic Risk Assessment in the Hands of Humans." *American Economic Journal: Economic Policy*, 16(4): 382–414.

A Proofs of the Main Results

A.1 Proofs for Section 2

Proof of Lemma 1. We provide one such model. Assume that, conditional on (C, Y^*), (H, T) has the following degenerate distribution:

$$Q((H,T) \in \cdot \mid C, Y^*) = \begin{cases} \delta_{(Y^*,1)}, & C = 1, \\ \delta_{(\mu_0,\mu_0/2)}, & C = 0, \end{cases} \tag{A1}$$

where $\mu_0 = P(Y^* = 1 \mid C = 0)$ and Q denotes the resulting joint law with (C, M, Y^*) marginal P. Note that

$$\begin{aligned} C = 1 &\implies H = Y^* \leq 1 = T, \\ C = 0 &\implies H = \mu_0 > \mu_0/2 = T. \end{aligned}$$

Hence, the threshold rule reproduces the choices. Since $\mu_0 \in (0, 1)$, T is a function of H. Consequently, $T \perp\!\!\!\perp (M, Y^*) \mid H$. Moreover,

$$H \in \{0, 1\} \implies Y^* = H$$

and

$$E_Q[Y^* \mid H = \mu_0] = E_P[Y^* \mid C = 0] = \mu_0.$$

Therefore, H is calibrated.

Proof of Proposition 2.1. If the decision maker's choices are consistent with Q, then P and Q have the same joint distribution of (M, C, Y^*). Summing over C gives (2). Note that the threshold rule and $T \perp\!\!\!\perp (M, Y^*) \mid H$ imply

$$\begin{aligned} Q(C = 1 \mid M, Y^*, H) &= Q(T \geq H \mid M, Y^*, H) \\ &= Q(T \geq H \mid H) = \rho_Q(H). \end{aligned}$$

Consequently, by iterated expectations,

$$Q(C = 1 \mid M = m, Y^* = y) = E_Q\big[\rho_Q(H) \mid M = m, Y^* = y\big],$$

which gives (3).

Conversely, assume that (2) and (3) hold. For every (m, y) with $P(M = m, Y^* = y) > 0$,

$$\begin{aligned} Q(M = m, C = 1, Y^* = y) &= Q(M = m, Y^* = y) E_Q\big[\rho_Q(H) \mid M = m, Y^* = y\big] \\ &= P(M = m, Y^* = y) P(C = 1 \mid M = m, Y^* = y) \\ &= P(M = m, C = 1, Y^* = y). \end{aligned}$$

Moreover, subtracting this equality from (2) gives

$$Q(M = m, C = 0, Y^* = y) = P(M = m, C = 0, Y^* = y).$$

Proof of Corollary 2.1. For any consistent threshold model, $T \perp\!\!\!\perp (M, Y^*) \mid H$ imply $C \perp\!\!\!\perp (M, Y^*) \mid H$. Calibration therefore gives $E_Q[Y^* \mid H, C] = H$. Consequently,

$$E_Q[H \mid C = c] = \mu_c, \qquad E_Q[H] = \mu = P(C = 1)\mu_1 + P(C = 0)\mu_0.$$

(i) Monotone choice behavior. Suppose first that $Q \in \mathcal{Q}^{\mathrm{MC}}(P)$. Since $E_Q[\rho_Q(H)] = P(C = 1)$ and $E_Q[H\rho_Q(H)] = E_Q[HC] = P(C = 1)\mu_1$,

$$\mathrm{Cov}_Q(H, \rho_Q(H)) = P(C = 1)P(C = 0)(\mu_1 - \mu_0).$$

Let H' be an independent copy of H. Monotonicity implies

$$2\,\mathrm{Cov}_Q(H, \rho_Q(H)) = E_Q\big[(H - H')(\rho_Q(H) - \rho_Q(H'))\big] \leq 0.$$

Hence, $\mu_1 \leq \mu_0$. If $\mu_1 = \mu_0$, the nonpositive integrand must be zero almost surely. It is strictly negative whenever $\rho_Q(H) \neq \rho_Q(H')$, so $\rho_Q(H) = \rho_Q(H')$ almost surely. By iterated expectations,

$$Q(C = 1 \mid M, Y^*) = E_Q[\rho_Q(H) \mid M, Y^*] = P(C = 1).$$

Therefore, $C \perp\!\!\!\perp (M, Y^*)$ under P.

Conversely, suppose $\mu_1 < \mu_0$. We provide one such model by extending P with

$$H = \mu_C, \qquad T = \frac{\mu_1 + \mu_0}{2}.$$

The threshold is positive and satisfies $\mu_1 < T < \mu_0$, so the threshold rule reproduces the choices. Since the two beliefs are distinct,

$$E_Q[Y^* \mid H = \mu_c] = E_P[Y^* \mid C = c] = \mu_c.$$

Hence, H is calibrated. The constant threshold satisfies conditional independence, and $\rho_Q(h) = \mathbf{1}\{h \leq T\}$ is nonincreasing, left-continuous, and equal to one at zero. Thus this model belongs to $\mathcal{Q}^{\mathrm{MC}}(P)$.

If instead $C \perp\!\!\!\perp (M, Y^*)$ under P, let $H = \mu$ and let T be independent of (M, Y^*) with distribution

$$Q(T \in \cdot) = P(C = 1)\delta_1 + P(C = 0)\delta_{\mu/2}.$$

It's straighforward to check that this model also belongs to $\mathcal{Q}^{\mathrm{MC}}(P)$.

(ii) Fixed threshold. Suppose $Q \in \mathcal{Q}^{\mathrm{FT}}(P)$ with $T = t$. The threshold rule gives $H \leq t$ when $C = 1$ and $H > t$ when $C = 0$. Consequently,

$$\mu_1 = E_Q[H \mid C = 1] \leq t < E_Q[H \mid C = 0] = \mu_0.$$

The second inequality is strict because $H - t$ is strictly positive conditional on $C = 0$. Thus $\mu_1 < \mu_0$ and $t \in [\mu_1, \mu_0)$. Conversely, if $\mu_1 < \mu_0$, the construction in (i) still applies.

A.2 Proofs for Section 4

Proof of Proposition 4.1. For any decision rule d based on H, calibration gives

$$S(d; Q) = \big(E_Q[Hd(H)], E_Q[d(H)]\big).$$

The realized score is

$$S(C; P) = \big(E_Q[H\rho_Q(H)], E_Q[\rho_Q(H)]\big).$$

Suppose first that the realized score lies on the information frontier. Choose $\tau \in [0, 1]$ such that

$$Q(H < \tau) \leq P(C = 1) \leq Q(H \leq \tau),$$

and let

$$d_\tau(h) = \mathbf{1}\{h < \tau\} + \alpha\mathbf{1}\{h = \tau\},$$

where $\alpha \in [0, 1]$ is chosen so that $E_Q[d_\tau(H)] = P(C = 1)$. Note that $(\rho_Q(H) - d_\tau(H))(H - \tau) \geq 0$ by construction. Since the two rules have the same selection rate and the realized score is on the frontier,

$$\begin{aligned} 0 &\leq E_Q\big[(\rho_Q(H) - d_\tau(H))(H - \tau)\big] \\ &= E_Q[H\rho_Q(H)] - E_Q[Hd_\tau(H)] \leq 0. \end{aligned}$$

Therefore, the integrand is zero almost surely. This implies $\rho_Q(H) = 1$ on $\{H < \tau\}$ and

$\rho_Q(H) = 0$ on $\{H > \tau\}$. Consequently, $C = \mathbf{1}\{H \leq \tau\}$ almost surely on $\{H \neq \tau\}$.

Conversely, if $\tau > 0$, for every decision rule d based on H, $E_Q\big[(\rho_Q(H) - d(H))(\tau - H)\big] \geq 0$. Thus the realized score maximizes $\tau S_2 - S_1$ over the feasible set and lies on the information frontier. If $\tau = 0$, the threshold models implies $C = 1$ on $\{H = 0\}$, while the cutoff decision rule implies $C = 0$ on $\{H > 0\}$. The realized score is therefore $(0, Q(H = 0))$, which also lies on the information frontier.

For Proposition 4.2, we start with showing the following lemma.

Lemma A1. *For any nonempty class $\mathcal{F}$ of threshold models consistent with P,*

$$\overline{\bigcup_{Q \in \mathcal{F}} \mathcal{S}(H; Q)} = \Pi \quad \Longleftrightarrow \quad \text{there are } Q_n \in \mathcal{F} \text{ with } E_{Q_n}[H(1-H)] \longrightarrow 0. \tag{A2}$$

Proof of Lemma A1. Notice that under calibration, $H(1-H) = \mathrm{Var}_Q(Y^* \mid H)$ measures the uncertainty about the outcome remaining after observing H and the score $(0, 1-\mu)$ corresponds to selecting all and only individuals with $Y^* = 0$.

If the closed union equals Π, there are rules $d_n(H)$ under models Q_n whose scores (s_{1n}, s_{2n}) converge to $(0, 1-\mu)$. By the definition of the score,

$$\begin{aligned} Q_n(C^{d_n} = 1, Y^* = 1) &= s_{1n} \longrightarrow 0, \\ Q_n(C^{d_n} = 0, Y^* = 0) &= 1 - \mu - Q_n(C^{d_n} = 1, Y^* = 0) \\ &= 1 - \mu - (s_{2n} - s_{1n}) \longrightarrow 0. \end{aligned}$$

Hence,

$$\begin{aligned} E_{Q_n}[H(1-H)] &\leq E_{Q_n}[H d_n(H) + (1-H)(1-d_n(H))] \\ &= s_{1n} + (1 - \mu - s_{2n} + s_{1n}) \longrightarrow 0, \end{aligned}$$

where the inequality follows because $H(1-H) \leq \min\{H, 1-H\}$.

Conversely, suppose $E_{Q_n}[H(1-H)] \to 0$. The decision rule $d(h) = 1 - h$ has score

$$S(d; Q_n) = \big(E_{Q_n}[H(1-H)], 1-\mu\big) \longrightarrow (0, 1-\mu).$$

The complementary rule $d(h) = h$ has limiting score (μ, μ), while the constant rules $d(h) = 0$ and $d(h) = 1$ yield $(0, 0)$ and $(\mu, 1)$, respectively. Randomizing among these four rules approximates every point of Π.

Proof of Proposition 4.2.

Baseline upper bound. We construct a sequence satisfying Lemma A1. Fix the joint distribution of (M, C, Y^*). For selected individuals, set $H = Y^*$. For $C = 0$, choose $0 < \varepsilon < \min\{\mu_0, 1 - \mu_0\}$ and let $H \in \{\varepsilon, 1 - \varepsilon\}$. Choose the conditional probabilities $q_y(h) := Q_\varepsilon(H = h \mid C = 0, Y^* = y)$ to satisfy

$$\sum_{h \in \{\varepsilon, 1-\varepsilon\}} q_y(h) = 1, \qquad y \in \{0, 1\},$$

$$\frac{\mu_0 q_1(h)}{\mu_0 q_1(h) + (1 - \mu_0) q_0(h)} = h, \qquad h \in \{\varepsilon, 1 - \varepsilon\}.$$

By Bayes' rule,

$$E_{Q_\varepsilon}[Y^* \mid C = 0, H = h] = h, \qquad h \in \{\varepsilon, 1 - \varepsilon\}.$$

This ensures $E_{Q_\varepsilon}[Y^* \mid H] = H$.

The selected and unselected beliefs have disjoint supports. Choose T as a function of H, equal to one at selected beliefs and strictly between zero and H at unselected beliefs. This reproduces the choices and ensures $T \perp\!\!\!\perp (M, Y^*) \mid H$. For the resulting model Q_ε,

$$E_{Q_\varepsilon}[H(1 - H)] = P(C = 0)\varepsilon(1 - \varepsilon) \longrightarrow 0.$$

Therefore, $\overline{\mathcal{S}}(P) = \Pi$.

(i) Monotone choice behavior. Suppose first that $\overline{\mathcal{S}}^{\mathrm{MC}}(P) = \Pi$. By (A2), there are models Q_n in this class with $E_{Q_n}[H(1 - H)] \to 0$. Calibration gives

$$E_{Q_n}[(1 - H)\mathbf{1}\{Y^* = 1\}] = E_{Q_n}[H\mathbf{1}\{Y^* = 0\}] = E_{Q_n}[H(1 - H)].$$

For every m with $P(M = m, Y^* = 1) > 0$,

$$\begin{aligned} Q_n(H \leq 1/2 \mid M = m, Y^* = 1) &= \frac{Q_n(H \leq 1/2, M = m, Y^* = 1)}{P(M = m, Y^* = 1)} \\ &\leq \frac{2E_{Q_n}[(1 - H)\mathbf{1}\{Y^* = 1\}]}{P(M = m, Y^* = 1)} \\ &= \frac{2E_{Q_n}[H(1 - H)]}{P(M = m, Y^* = 1)} \longrightarrow 0. \end{aligned}$$

Similarly, for every m with $P(M = m, Y^* = 0) > 0$,

$$Q_n(H > 1/2 \mid M = m, Y^* = 0) \longrightarrow 0.$$

Conditional independence gives $\alpha_y(m) = E_{Q_n}[\rho_{Q_n}(H) \mid M = m, Y^* = y], \forall n$. Since ρ_{Q_n} is

nonincreasing,

$$\alpha_1(m) - \alpha_0(m') \le Q_n(H \le 1/2 \mid M = m, Y^* = 1) \\ + Q_n(H > 1/2 \mid M = m', Y^* = 0).$$

Taking limits shows there is no reversal across the algorithm's predictions.

Conversely, suppose there is no reversal across the algorithm's predictions. Choose $\max \alpha_1(m) \le z \le \min \alpha_0(m)$. The assumptions $P(C = 1, Y^* = 1) > 0$ and $P(C = 0, Y^* = 0) > 0$ imply $0 < z < 1$. For $0 < \varepsilon < 1/2$, let $\mathcal{H}_\varepsilon = \{0, \varepsilon, 1 - \varepsilon, 1\}$ and set

$$\rho_{Q_\varepsilon}(0) = 1, \qquad \rho_{Q_\varepsilon}(\varepsilon) = \rho_{Q_\varepsilon}(1 - \varepsilon) = z, \qquad \rho_{Q_\varepsilon}(1) = 0.$$

Fix the observed marginal of (M, Y^*). Choose the conditional probabilities $q_y(h \mid m) := Q_\varepsilon(H = h \mid M = m, Y^* = y)$ to satisfy

$$\sum_{h \in \mathcal{H}_\varepsilon} q_y(h \mid m) = 1, \qquad q_1(0 \mid m) = q_0(1 \mid m) = 0,$$

$$q_y(0 \mid m) + z\big[q_y(\varepsilon \mid m) + q_y(1 - \varepsilon \mid m)\big] = \alpha_y(m).$$

The last equation matches the observed choice probability. Combining it with normalization gives

$$q_0(\varepsilon \mid m) + q_0(1 - \varepsilon \mid m) = \frac{1 - \alpha_0(m)}{1 - z} \in [0, 1],$$
$$q_1(\varepsilon \mid m) + q_1(1 - \varepsilon \mid m) = \frac{\alpha_1(m)}{z} \in [0, 1],$$

where the bounds follow from $\alpha_1(m) \le z \le \alpha_0(m)$. The remaining probability is assigned to $H = 0$ when $Y^* = 0$ and to $H = 1$ when $Y^* = 1$. Averaging over M gives

$$Q_\varepsilon(0 < H < 1, Y^* = 0) = \frac{P(C = 0, Y^* = 0)}{1 - z} > 0,$$
$$Q_\varepsilon(0 < H < 1, Y^* = 1) = \frac{P(C = 1, Y^* = 1)}{z} > 0.$$

Since the two masses above are fixed and positive, choose ε sufficiently small that

$$\varepsilon < \frac{Q_\varepsilon(0 < H < 1, Y^* = 1)}{Q_\varepsilon(0 < H < 1)} < 1 - \varepsilon.$$

Choose $Q_\varepsilon(H=\varepsilon)$ and $Q_\varepsilon(H=1-\varepsilon)$ to satisfy

$$\begin{aligned} Q_\varepsilon(H=\varepsilon)+Q_\varepsilon(H=1-\varepsilon) &= Q_\varepsilon(0<H<1),\\ \varepsilon Q_\varepsilon(H=\varepsilon)+(1-\varepsilon)Q_\varepsilon(H=1-\varepsilon) &= Q_\varepsilon(0<H<1,Y^*=1). \end{aligned}$$

The preceding inequality ensures that both probabilities are positive. For $h\in\{\varepsilon,1-\varepsilon\}$, set

$$\begin{aligned} q_0(h\mid m) &= \frac{1-\alpha_0(m)}{1-z}\frac{(1-h)Q_\varepsilon(H=h)}{Q_\varepsilon(0<H<1,Y^*=0)},\\ q_1(h\mid m) &= \frac{\alpha_1(m)}{z}\frac{hQ_\varepsilon(H=h)}{Q_\varepsilon(0<H<1,Y^*=1)}. \end{aligned}$$

These assignments preserve the previously specified sum $q_y(\varepsilon\mid m)+q_y(1-\varepsilon\mid m)$ in each cell and, for $h\in\{\varepsilon,1-\varepsilon\}$, give

$$\begin{aligned} \sum_m P(M=m,Y^*=0)q_0(h\mid m) &= (1-h)Q_\varepsilon(H=h),\\ \sum_m P(M=m,Y^*=1)q_1(h\mid m) &= hQ_\varepsilon(H=h). \end{aligned}$$

Thus $E_{Q_\varepsilon}[Y^*\mid H=h]=h$ at these two beliefs; $q_1(0\mid m)=q_0(1\mid m)=0$ gives calibration at zero and one. Since both beliefs ε and $1-\varepsilon$ have choice probability z, the observed choice probabilities are preserved.

Choose T independently of (M,Y^*,H) with

$$Q_\varepsilon(T=t_L)=1-z,\qquad Q_\varepsilon(T=t_U)=z,\qquad 0<t_L<\varepsilon<1-\varepsilon\le t_U<1.$$

Then $C=\mathbf{1}\{H\le T\}$ implements the displayed choice probabilities, giving $Q_\varepsilon\in\mathcal{Q}^{\mathrm{MC}}(P)$. Moreover,

$$E_{Q_\varepsilon}[H(1-H)]=Q_\varepsilon(0<H<1)\varepsilon(1-\varepsilon)\le\varepsilon(1-\varepsilon)\longrightarrow 0.$$

Hence, $\overline{\mathcal{S}}^{\mathrm{MC}}(P)=\Pi$.

(ii) Fixed threshold. For any consistent fixed-threshold model, Corollary 2.1 gives $\mu_1\le T<\mu_0$. Since $H\le T$ when $C=1$,

$$\begin{aligned} E_Q[H(1-H)] &\ge E_Q[H(1-H)C]\\ &\ge (1-\mu_0)E_Q[HC]\\ &= (1-\mu_0)P(C=1,Y^*=1)>0. \end{aligned}$$

This positive lower bound is the same for every model in $\mathcal{Q}^{\mathrm{FT}}(P)$. Hence, Lemma A1 rules out equality with Π.

Lower bound. For every consistent model, $\mathcal{S}(C;P) \subseteq \mathcal{S}(H;Q)$. Choose $\varepsilon > 0$ so that $\mu_0 \pm \varepsilon \in (0,1)$ and both values differ from μ_1. Extend P with $H = \mu_1$ when $C = 1$ and, for $h \in \{\mu_0 - \varepsilon, \mu_0 + \varepsilon\}$,

$$Q_\varepsilon(H = h \mid C = 0, Y^* = y) = \begin{cases} h/(2\mu_0), & y = 1, \\ (1-h)/(2(1-\mu_0)), & y = 0. \end{cases}$$

Then,

$$E_{Q_\varepsilon}[Y^* \mid C = 0, H = h] = \frac{\mu_0\, h/(2\mu_0)}{1/2} = h.$$

Thus H is calibrated. Since H determines C, setting $T = 1$ when $C = 1$ and $T = H/2$ when $C = 0$ gives a consistent threshold model.

For any rule $d(H)$, the rule $c \mapsto E_{Q_\varepsilon}[d(H) \mid C = c]$ under P has the same selection rate as $d(H)$ and first score coordinate $E_{Q_\varepsilon}[\mu_C d(H)]$. Since $|H - \mu_C| \le \varepsilon$,

$$\left|E_{Q_\varepsilon}[Hd(H)] - E_{Q_\varepsilon}[\mu_C d(H)]\right| = \left|E_{Q_\varepsilon}[(H - \mu_C)d(H)]\right| \le \varepsilon.$$

Thus every score in $\mathcal{S}(H;Q_\varepsilon)$ is within ε of $\mathcal{S}(C;P)$. Letting $\varepsilon \downarrow 0$ and using that $\mathcal{S}(C;P)$ is closed gives $\underline{\mathcal{S}}(P) = \mathcal{S}(C;P)$.

For each nonempty restricted class, the constructions in Corollary 2.1 attain $\mathcal{S}(H;Q) = \mathcal{S}(C;P)$. Hence both restricted lower bounds equal $\mathcal{S}(C;P)$.

A.3 Proofs for Sections 5 and 6

Proof of Lemma 2.

(i) If a consistent model satisfies $H \perp\!\!\!\perp Y^* \mid M = m$, then conditional independence gives

$$\begin{aligned} P(C = 1 \mid M = m, Y^*) &= E_Q[\rho_Q(H) \mid M = m, Y^*] \\ &= E_Q[\rho_Q(H) \mid M = m]. \end{aligned}$$

Thus $C \perp\!\!\!\perp Y^* \mid M = m$.

Conversely, suppose $C \perp\!\!\!\perp Y^* \mid M = m$ and choose another prediction $m' \neq m$. Fix the joint distribution of (M, C, Y^*) and set $H = \mu_1$ when $C = 1$. When $C = 0$, assign beliefs

$\mu_0 \pm \varepsilon$ by

$$\begin{aligned} &Q_\varepsilon(M = j, C = 0, Y^* = y, H = \mu_0 \pm \varepsilon) \\ &\quad = \frac{P(M = j, C = 0, Y^* = y)}{2} \pm \frac{(2y-1)P(C=0)\varepsilon}{2}\mathbf{1}\{j = m'\}. \end{aligned}$$

The condition $\min_{y\in\{0,1\}} P(M = m', C = 0, Y^* = y) > 0$ allows $\varepsilon > 0$ small enough that these masses are nonnegative, $\mu_0 \pm \varepsilon \in (0, 1)$, and both beliefs differ from μ_1. Summing over the two beliefs preserves P, and

$$Q_\varepsilon(H = \mu_0 \pm \varepsilon) = \frac{P(C=0)}{2}, \qquad Q_\varepsilon(Y^* = 1, H = \mu_0 \pm \varepsilon) = \frac{P(C=0)(\mu_0 \pm \varepsilon)}{2}.$$

Hence H is calibrated. Set $T = 1$ when $H = \mu_1$ and $T = H/2$ when $H = \mu_0 \pm \varepsilon$. Since T is a function of H, the resulting model satisfies the threshold rule and conditional independence. At the target prediction $m \neq m'$,

$$\begin{aligned} Q_\varepsilon(H = \mu_1 \mid M = m, Y^* = y) &= P(C = 1 \mid M = m), \\ Q_\varepsilon(H = \mu_0 \pm \varepsilon \mid M = m, Y^* = y) &= \frac{P(C = 0 \mid M = m)}{2}. \end{aligned}$$

These probabilities do not depend on y, so $H \perp\!\!\!\perp Y^* \mid M = m$ in this consistent model.

(ii) The value of observing the choice. For any $\lambda > 0$,

$$\begin{aligned} V_\lambda((M, C); P) &= \sum_{m,c} P(M = m, C = c)(1 - \lambda\mu_c(m))_+, \\ V_\lambda(M; P) &= \sum_{m} P(M = m)(1 - \lambda\mu(m))_+. \end{aligned}$$

Since $\mu(m) = E_P[\mu_C(m) \mid M = m]$, convexity gives the weak inequality. It is strict exactly when some m satisfies

$$\min\{\mu_0(m), \mu_1(m)\} < \frac{1}{\lambda} < \max\{\mu_0(m), \mu_1(m)\}.$$

Such a $\lambda > 0$ exists if and only if $\mu_0(m) \neq \mu_1(m)$ for some m, equivalently $C \not\perp\!\!\!\perp Y^* \mid M = m$. Part (i) then gives the result.

Proof of Lemma 3. Write $p_{cy}(m) = P(C = c, M = m, Y^* = y)$ and $\mu_Q(M, H) = Q(Y^* = 1 \mid M, H)$. Since $u(1 - u) \leq \min\{u, 1 - u\} \leq 2u(1 - u)$ for $u \in [0, 1]$, Definition 4 implies

that t is revealing if and only if there exists a sequence of models $Q_n \in \mathcal{Q}^{\mathrm{FT}}(P)$ satisfying

$$Q_n(T = t) = 1 \quad \text{for every } n, \qquad \eta_n := E_{Q_n}[\mu_{Q_n}(M,H)(1-\mu_{Q_n}(M,H))] \longrightarrow 0.$$

Note that $\mathcal{Q}^{\mathrm{FT}}(P) \neq \emptyset$ gives $\mu_1 < \mu_0$.

Suppose first that t is revealing. Corollary 2.1 gives $\mu_1 \leq t < \mu_0$. If $t = \mu_1$, then $H \leq t$ on $\{C = 1\}$ and $E_Q[H \mid C = 1] = \mu_1 = t$ imply $H = t$ on $\{C = 1\}$. Hence every consistent model with $T = t$ satisfies

$$E_Q[\min\{\mu_Q(M,H), 1-\mu_Q(M,H)\}] \geq \sum_m \min\{p_{10}(m), p_{11}(m)\} > 0,$$

contradicting the definition. Thus $\mu_1 < t < \mu_0$.

For the sequence above, define

$$\mathcal{H}_n^1(m) := \{h : \mu_{Q_n}(m,h) > 1/2\}, \qquad \mathcal{H}_n^0(m) := \{h : \mu_{Q_n}(m,h) \leq 1/2\}.$$

For a given algorithm prediction m, $\mathcal{H}_n^1(m)$ contains the belief values h for which $Y^* = 1$ is more likely than $Y^* = 0$ conditional on $(M,H) = (m,h)$, with the analogous interpretation for $\mathcal{H}_n^0(m)$.

Notice that

$$\begin{aligned} Q_n(M = m, Y^* = 1, H \in \mathcal{H}_n^0(m)) &= E_{Q_n}[\mu_{Q_n}(M,H)\mathbf{1}\{M = m, H \in \mathcal{H}_n^0(m)\}] \\ &\leq 2\eta_n \longrightarrow 0, \\ Q_n(M = m, Y^* = 0, H \in \mathcal{H}_n^1(m)) &= E_{Q_n}[(1-\mu_{Q_n}(M,H))\mathbf{1}\{M = m, H \in \mathcal{H}_n^1(m)\}] \\ &\leq 2\eta_n \longrightarrow 0. \end{aligned} \tag{A3}$$

The threshold rule gives $H > t$ for unselected individuals and $H \leq t$ for selected individuals. Hence,

$$\begin{aligned} H > t &\implies H \geq \frac{t}{1-t}(1-H), \\ H \leq t &\implies 1 - H \geq \frac{1-t}{t}H. \end{aligned}$$

Taking expectations on $\mathcal{H}_n^0(m) \cap (t,1]$ and $\mathcal{H}_n^1(m) \cap [0,t]$ and using calibration gives

$$\begin{aligned} Q_n(C=0, Y^*=1, H \in \mathcal{H}_n^0(m)) &= E_{Q_n}[H\mathbf{1}\{H \in \mathcal{H}_n^0(m) \cap (t,1]\}] \\ &\geq \frac{t}{1-t} Q_n(C=0, Y^*=0, H \in \mathcal{H}_n^0(m)), \\ Q_n(C=1, Y^*=0, H \in \mathcal{H}_n^1(m)) &= E_{Q_n}[(1-H)\mathbf{1}\{H \in \mathcal{H}_n^1(m) \cap [0,t]\}] \\ &\geq \frac{1-t}{t} Q_n(C=1, Y^*=1, H \in \mathcal{H}_n^1(m)). \end{aligned}$$

For $C=1$, splitting the event according to the two belief sets gives

$$\begin{aligned} Q_n(C=1, Y^*=1, H \in \mathcal{H}_n^1(m)) &\geq Q_n(M=m, C=1, Y^*=1, H \in \mathcal{H}_n^1(m)) \\ &= p_{11}(m) - Q_n(M=m, C=1, Y^*=1, H \in \mathcal{H}_n^0(m)) \\ &\geq p_{11}(m) - 2\eta_n. \end{aligned}$$

For the opposite outcome, summing over the algorithm's predictions gives

$$\begin{aligned} Q_n(C=1, Y^*=0, H \in \mathcal{H}_n^1(m)) &= \sum_{m'} Q_n(M=m', C=1, Y^*=0, H \in \mathcal{H}_n^1(m)) \\ &\leq 2\eta_n + \sum_{m' \neq m} p_{10}(m'). \end{aligned}$$

The same steps for $C=0$ give

$$\begin{aligned} Q_n(C=0, Y^*=0, H \in \mathcal{H}_n^0(m)) &\geq p_{00}(m) - 2\eta_n, \\ Q_n(C=0, Y^*=1, H \in \mathcal{H}_n^0(m)) &\leq 2\eta_n + \sum_{m' \neq m} p_{01}(m'). \end{aligned}$$

Together with the inequalities above, these give

$$\begin{aligned} \frac{1-t}{t}\big(p_{11}(m) - 2\eta_n\big) &\leq Q_n(C=1, Y^*=0, H \in \mathcal{H}_n^1(m)) \\ &\leq \sum_{m' \neq m} p_{10}(m') + 2\eta_n, \\ \frac{t}{1-t}\big(p_{00}(m) - 2\eta_n\big) &\leq Q_n(C=0, Y^*=1, H \in \mathcal{H}_n^0(m)) \\ &\leq \sum_{m' \neq m} p_{01}(m') + 2\eta_n. \end{aligned} \tag{A4}$$

Taking limits and rearranging gives $\underline{T} \leq t \leq \overline{T}$.

Conversely, fix $t \in [\underline{T}, \overline{T}] \cap (\mu_1, \mu_0)$. We first form groups in which M identifies Y^*. For

$m \neq m'$, let $q_c(m, m')$ be the total probability of a group with choice c, containing observations with $(M, Y^*) = (m, 0)$ and $(M, Y^*) = (m', 1)$. Assign probabilities $(1-t)q_c(m, m')$ and $tq_c(m, m')$ to these two types, respectively. The group's outcome mean is then t, and its two outcomes are distinguished by M.

To match the observed probabilities, choose $q_c(m, m') \geq 0$, with $q_c(m, m) = 0$, satisfying

$$t \sum_{m''} q_1(m'', m') = p_{11}(m'), \qquad (1-t) \sum_{m''} q_1(m, m'') \leq p_{10}(m),$$
$$(1-t) \sum_{m''} q_0(m, m'') = p_{00}(m), \qquad t \sum_{m''} q_0(m'', m') \leq p_{01}(m').$$

For $C = 1$, the observations counted by $p_{11}(m')$ make up fraction t of their groups. Those groups therefore have total probability $p_{11}(m')/t$. The remaining fraction $1-t$ contributes probability $(1-t)p_{11}(m')/t$ with $Y^* = 0$, which must come from predictions $m \neq m'$. Since $t \geq \underline{T}$, the required probability is available for each m' and in total:

$$\frac{1-t}{t} p_{11}(m') \leq \sum_{m \neq m'} p_{10}(m),$$
$$\frac{1-t}{t} \sum_{m'} p_{11}(m') \leq \sum_{m} p_{10}(m).$$

For $C = 0$, the observations counted by $p_{00}(m)$ instead make up fraction $1-t$ of their groups, so the probability needed with $Y^* = 1$ is $tp_{00}(m)/(1-t)$. Since $t \leq \overline{T}$,

$$\frac{t}{1-t} p_{00}(m) \leq \sum_{m' \neq m} p_{01}(m'),$$
$$\frac{t}{1-t} \sum_{m} p_{00}(m) \leq \sum_{m'} p_{01}(m').$$

These inequalities also allow the groups to be formed simultaneously. For two distinct predictions m' and m'', every m differs from at least one of them. Thus a collection of at least two predictions can use all observations with the opposite outcome, and its requirement is covered by the total inequality.

The equalities for q_1 and q_0 assign all observations with $C = 1, Y^* = 1$ and $C = 0, Y^* = 0$

to groups. The only probabilities left unassigned are

$$p_{10}(m) - (1-t)\sum_{m''} q_1(m, m'') \geq 0,$$
$$p_{01}(m') - t\sum_{m''} q_0(m'', m') \geq 0.$$

Put all remaining observations with $C = 1$ in one group, whose outcome is always zero, and those with $C = 0$ in another, whose outcome is always one.

Let G record the groups with positive probability and let Q_0 denote the resulting distribution of (M, C, Y^*, G). It has (M, C, Y^*) marginal P, and Y^* is determined by (M, G). Write c_g for the choice in group g and $h_g = Q_0(Y^* = 1 \mid G = g)$ for its outcome mean.

For each group $G = g$, assign $H = h_g$. For H to identify the group and reproduce the choices, we need distinct h_g's on the appropriate sides of t. Keep the group probabilities fixed and, for $0 < \varepsilon < 1$, replace the conditional distribution of (M, Y^*) in group g by

$$(1-\varepsilon)Q_0(M, Y^* \mid G = g) + \varepsilon P(M, Y^* \mid C = c_g).$$

Averaging over groups preserves P, while the expected value of Y^* of each group becomes

$$(1-\varepsilon)h_g + \varepsilon\mu_{c_g} \begin{cases} < t, & c_g = 1, \\ > t, & c_g = 0, \end{cases}$$

because $h_g \leq t$ for selected groups, $h_g \geq t$ for unselected groups, and $\mu_1 < t < \mu_0$.

Any means that still coincide can be separated by exchanging equal probabilities of $Y^* = 0$ and $Y^* = 1$ in opposite directions between groups with the same choice, at a fixed prediction. Positivity permits such exchanges to vanish with ε while retaining the strict inequalities above.

Write Q_ε for the resulting distribution, set $H = E_{Q_\varepsilon}[Y^* \mid G]$, and set $T \equiv t$. Then H identifies G and is calibrated, and $C = \mathbf{1}\{H \leq t\}$. Moreover,

$$E_{Q_\varepsilon}[\min\{\mu_{Q_\varepsilon}(M, H), 1 - \mu_{Q_\varepsilon}(M, H)\}]$$
$$= \sum_{g,m} \min_{y \in \{0,1\}} Q_\varepsilon(G = g, M = m, Y^* = y) \longrightarrow 0,$$

since under Q_0 each (M, G) has only one possible outcome. Thus t is revealing.

Remark A1 (Limits of thresholds)**.** Suppose $\mu_1 < \mu_0$. A value t belongs to $[\underline{T}, \overline{T}]$ if and

only if there are models $Q_n \in \mathcal{Q}^{\mathrm{FT}}(P)$ with constant thresholds t_n satisfying

$$t_n \longrightarrow t, \qquad \eta_n := E_{Q_n}[\mu_{Q_n}(M,H)(1-\mu_{Q_n}(M,H))] \longrightarrow 0.$$

For any $t \in [\underline{T}, \overline{T}]$, the construction in the proof of Lemma 3 works with

$$T = t_\varepsilon = (1-\varepsilon)t + \varepsilon\frac{\mu_1+\mu_0}{2}.$$

Before the exchanges, selected means are at most $(1-\varepsilon)t + \varepsilon\mu_1$, and unselected means are at least $(1-\varepsilon)t + \varepsilon\mu_0$. Since $\mu_1 < \mu_0$,

$$(1-\varepsilon)t + \varepsilon\mu_1 < t_\varepsilon < (1-\varepsilon)t + \varepsilon\mu_0. \tag{A5}$$

The strict inequalities in (A5) hold even when $t = \mu_1$ or $t = \mu_0$. Choose the exchanges small enough to keep the group means on their respective sides of t_ε. Then $t_\varepsilon \to t$, and the calculation in that proof together with $u(1-u) \le \min\{u, 1-u\}$ gives the required convergence. Conversely, the inequalities in (A4) apply with t_n in place of t. Since $\mu_1 \le t_n < \mu_0$, $t_n \to t$ and $\eta_n \to 0$ imply $t \in [\underline{T}, \overline{T}]$.

Proof of Proposition 5.1. Write $\mu_Q(M,H) = Q(Y^* = 1 \mid M, H)$ and $\eta_Q = E_Q[\mu_Q(M,H)(1-\mu_Q(M,H))]$ as in the proof of Lemma 3. Following the definitions, the loss relative to perfect prediction is

$$1 - \mu - V_\lambda((M,H);Q) = E_Q\big[\min\{(\lambda-1)\mu_Q(M,H), 1-\mu_Q(M,H)\}\big].$$

For $s \in [0,1]$,

$$\min\{1, \lambda-1\}s(1-s) \le \min\{(\lambda-1)s, 1-s\} \le \lambda s(1-s).$$

Consequently,

$$\min\{1, \lambda-1\}\eta_Q \le 1 - \mu - V_\lambda((M,H);Q) \le \lambda\eta_Q.$$

Thus the upper bound over any nonempty class equals $1-\mu$ if and only if that class contains a sequence with $\eta_Q \to 0$.

Baseline upper bound. By Proposition 4.2, H, and hence (M,H), can approach perfect prediction, giving $\overline{V}_\lambda(P) = 1-\mu$.

(i) Monotone choice behavior. If there is no reversal across the algorithm's predictions, Proposition 4.2 shows that H, and hence (M,H), can approach perfect prediction under

monotone choice behavior, giving $\overline{V}_\lambda^{\mathrm{MC}}(P) = 1-\mu$. Otherwise, Corollary 2.1 gives $\mu_1 < \mu_0$. If $\underline{T} \le \overline{T}$, Remark A1 then supplies a sequence with $\eta_Q \to 0$ in $\mathcal{Q}^{\mathrm{FT}}(P) \subseteq \mathcal{Q}^{\mathrm{MC}}(P)$, giving the upper bound $1-\mu$.

To prove the strict inequality under an excess reversal, fix a prediction m and a model $Q \in \mathcal{Q}^{\mathrm{MC}}(P)$. Write $p_y(m) = P(M = m, Y^* = y)$ and $F_y(h) = Q(H \le h \mid M = m, Y^* = y)$. Use the sets $\mathcal{H}_Q^1(m)$ and $\mathcal{H}_Q^0(m)$ as in the proof of Lemma 3, with Q in place of Q_n. For $0 < h < 1$, the corresponding $2\eta_Q$ bounds in (A3) and calibration on $\mathcal{H}_Q^1(m) \cap [0,h]$ and $\mathcal{H}_Q^0(m) \cap (h,1]$ give, respectively,

$$\begin{aligned} p_1(m)F_1(h) &= Q(M = m, Y^* = 1, H \le h) \\ &\le \frac{h}{1-h}\big(P(M \ne m, Y^* = 0) + 2\eta_Q\big) + 2\eta_Q, \\ p_0(m)(1 - F_0(h)) &= Q(M = m, Y^* = 0, H > h) \\ &\le \frac{1-h}{h}\big(P(M \ne m, Y^* = 1) + 2\eta_Q\big) + 2\eta_Q. \end{aligned}$$

Using $F_y \in [0,1]$ and $\min\{1,a\} + \min\{1,b\} - 1 \le ab$ for $a, b \ge 0$, we obtain

$$\begin{aligned} F_1(h) - F_0(h) \le{}& \frac{\big(P(M \ne m, Y^* = 0) + 2\eta_Q\big)\big(P(M \ne m, Y^* = 1) + 2\eta_Q\big)}{p_1(m)p_0(m)} \\ &+ 2\eta_Q\left(\frac{1}{p_1(m)} + \frac{1}{p_0(m)}\right). \end{aligned}$$

Since ρ_Q is nonincreasing, each event $\{\rho_Q(H) > u\}$ has the form $\{H \le h\}$ or $\{H < h\}$, possibly empty. Also,

$$\alpha_y(m) = \int_0^1 Q(\rho_Q(H) > u \mid M = m, Y^* = y)\, du.$$

Averaging these differences gives

$$\alpha_1(m) - \alpha_0(m) \le \sup_{h \in [0,1]} \big(F_1(h) - F_0(h)\big).$$

If $\eta_{Q_n} \to 0$ along any sequence in this class, the preceding bounds imply

$$\alpha_1(m) - \alpha_0(m) \le \frac{P(M \ne m, Y^* = 0)P(M \ne m, Y^* = 1)}{p_1(m)p_0(m)}.$$

An excess reversal contradicts this inequality, so the upper bound is strictly below $1-\mu$.

(ii) A fixed threshold. By Remark A1 and compactness of $[\mu_1, \mu_0]$, the required sequence

of fixed-threshold models with $\eta_Q \to 0$ exists if and only if $\underline{T} \le \overline{T}$.

Lower bound. For each nonempty class, the lower-bound constructions in the proof of Proposition 4.2 give models Q_ε satisfying

$$E_{Q_\varepsilon}\big[|\mu_{Q_\varepsilon}(M,H) - \mu_C(M)|\big] \longrightarrow 0.$$

Since $|(1-\lambda u)_+ - (1-\lambda v)_+| \le \lambda |u-v|$,

$$\big|V_\lambda((M,H);Q_\varepsilon) - V_\lambda((M,C);P)\big| \le \lambda E_{Q_\varepsilon}\big[|\mu_{Q_\varepsilon}(M,H) - \mu_C(M)|\big] \longrightarrow 0.$$

Thus all three lower bounds equal $V_\lambda((M,C);P)$.

Proof of Proposition 6.1. For a given threshold model Q, the value of AI assistance is

$$\Gamma(Q) := E_Q\big[(1-\mu_Q(M,H)/T)_+ - (1-H/T)_+\big].$$

Since $0 < T \le 1$, conditional Jensen and $(1-\mu_Q(M,H)/T)_+ \le 1-\mu_Q(M,H)$ give

$$0 \le \Gamma(Q) \le 1 - \mu - E_Q[(1-H/T)_+] \le 1-\mu.$$

Baseline upper bound. Fix a permutation σ of $\mathcal{M}$ with $\sigma(m) \neq m$ for every m. For each (c,m), pair the mass $p_{c1}(m)$ with $p_{c0}(\sigma(m))$. Choose pairwise distinct beliefs $h^-_{cm}, h^+_{cm} \in (0,1)$ satisfying

$$\frac{h^-_{cm} + h^+_{cm}}{2} = \frac{p_{c1}(m)}{p_{c1}(m) + p_{c0}(\sigma(m))}.$$

For $h \in \{h^-_{cm}, h^+_{cm}\}$, assign

$$\begin{aligned} Q(M=m, C=c, Y^*=1, H=h) &= \frac{h}{2}\big[p_{c1}(m) + p_{c0}(\sigma(m))\big], \\ Q(M=\sigma(m), C=c, Y^*=0, H=h) &= \frac{1-h}{2}\big[p_{c1}(m) + p_{c0}(\sigma(m))\big]. \end{aligned}$$

Summing over the two beliefs recovers the original masses, and the beliefs are calibrated. Since $m \neq \sigma(m)$ and all beliefs are distinct, (M,H) reveals Y^* and H determines C. Set

$$T = \begin{cases} H, & C = 1, \\ H/2, & C = 0. \end{cases}$$

Then $T > 0$ is a function of H, the threshold rule reproduces C, and

$$E_Q[(1 - H/T)_+] = 0, \qquad E_Q[(1 - \mu_Q(M, H)/T)_+] = P(Y^* = 0) = 1 - \mu.$$

Hence $\Gamma(Q) = 1 - \mu$, attaining the baseline upper bound.

(i) Monotone-choice upper bound. If $\mu_1 < \mu_0$ and $\mu_1 \in [\underline{T}, \overline{T}]$, the construction in Remark A1 provides fixed-threshold models with thresholds approaching μ_1 and posteriors approaching perfect prediction. For any consistent fixed-threshold sequence Q_n with thresholds t_n,

$$1 - \mu - \Gamma(Q_n) = 1 - \mu - V_{1/t_n}((M, H); Q_n) + P(C = 1)\left(1 - \frac{\mu_1}{t_n}\right). \tag{A6}$$

Since $\mu_1 \leq t_n < \mu_0$,

$$\begin{aligned} &\min\left\{1, \frac{1 - \mu_0}{\mu_0}\right\} E_{Q_n}[\mu_{Q_n}(M, H)(1 - \mu_{Q_n}(M, H))] \\ &\leq 1 - \mu - V_{1/t_n}((M, H); Q_n) \\ &\leq \frac{1}{\mu_1} E_{Q_n}[\mu_{Q_n}(M, H)(1 - \mu_{Q_n}(M, H))]. \end{aligned} \tag{A7}$$

Hence $\Gamma(Q_n) \to 1 - \mu$. Since these models also satisfy monotone choice behavior, its upper bound is $1 - \mu$ as well.

If $\mu_1 = \mu_0$, nonemptiness implies $C \perp\!\!\!\perp (M, Y^*)$. Keep the joint distribution of (M, Y^*, H) from the baseline upper-bound construction, where $H > 0$ is calibrated and (M, H) reveals Y^*. Choose C independently of (M, Y^*, H) with $Q(C = 1) = P(C = 1)$. Set $T = H$ when $C = 1$ and $T = H/2$ otherwise. The choice probability has the nonincreasing, left-continuous version

$$\rho_Q(h) = \begin{cases} 1, & h = 0, \\ P(C = 1), & h > 0. \end{cases}$$

Finally, $H/T \in \{1, 2\}$ and $\mu_Q(M, H) = Y^*$ give

$$(1 - H/T)_+ = 0, \qquad (1 - \mu_Q(M, H)/T)_+ = 1 - Y^*.$$

Hence $\Gamma(Q) = E_Q[1 - Y^*] = 1 - \mu$.

(ii) Fixed-threshold upper bound. The fixed-threshold construction in part (i) proves sufficiency; necessity follows from (A6), (A7), and Remark A1.

Lower bounds. When $\mu_1 \neq \mu_0$, set $H = \mu_C$ and take

$$T = \begin{cases} 1, & C = 1, \\ \delta, & C = 0, \end{cases} \qquad 0 < \delta < \min\{\mu_0, \min_m \mu_0(m)\}.$$

This is a consistent calibrated threshold model. The optimal rule selects when $T = 1$ and does not select when $T = \delta$, so $\Gamma(Q) = 0$. Its choice probability is nonincreasing when $\mu_1 < \mu_0$.

For the baseline class with $\mu_1 = \mu_0$, use the lower-bound construction in Proposition 4.2, with $H = \mu_0$ when $C = 1$ and $H \in \{\mu_0 - \varepsilon, \mu_0 + \varepsilon\}$ when $C = 0$. These beliefs determine C, and their finitely many posterior values are positive. Set $T = 1$ when $C = 1$ and $T = \delta$ when $C = 0$, choosing $\delta > 0$ so that

$$\delta < \min\{H, \mu_Q(M, H)\} \quad \text{on } \{C = 0\}.$$

In the remaining nonempty monotone case, $C \perp\!\!\!\perp (M, Y^*)$. Set $H = \mu$ and choose T independently of (M, Y^*) with

$$Q(T = 1) = P(C = 1), \qquad Q(T = \delta) = P(C = 0),$$

where $0 < \delta < \min_m P(Y^* = 1 \mid M = m) \leq \mu$. Set $C = \mathbf{1}\{H \leq T\}$; the constant belief gives monotone choice behavior. In both constructions, the optimal rules using H and (M, H) select when $T = 1$ and do not select when $T = \delta$, so $\Gamma(Q) = 0$.

For a fixed threshold $t \in [\mu_1, \mu_0)$, Proposition 5.1 gives

$$\begin{aligned} \Gamma(Q) &= V_{1/t}((M, H); Q) - V_{1/t}(C; P) \\ &\geq V_{1/t}((M, C); P) - V_{1/t}(C; P). \end{aligned}$$

Equality holds for $H = \mu_C$, $T = t$. The value difference is continuous in t, and its value at μ_0 is approached as $t \uparrow \mu_0$ through feasible thresholds. Thus its infimum over $[\mu_1, \mu_0)$ equals its minimum over $[\mu_1, \mu_0]$. Moreover,

$$\begin{aligned} t\big(V_{1/t}((M, C); P) - V_{1/t}(C; P)\big) = &\sum_m P(M = m, C = 1)(\mu_1(m) - t)_+ \\ &+ \sum_m P(M = m, C = 0)(t - \mu_0(m))_+. \end{aligned}$$

This vanishes for some $t \in [\mu_1, \mu_0]$ exactly when $\max_m \mu_1(m) \leq \min_m \mu_0(m)$.

B Additional Results for the Threshold Model

B.1 Microfoundation

This appendix gives the expected utility interpretation of the threshold model behind Remark 2.1 and the normalization behind Definition 1. In this appendix, we now interpret H as the decision maker's *belief* (their subjective probability that the outcome occurs) and T is their *loss ratio*. In the weakest model of a decision maker who forms a belief and a loss ratio from what they observe and from shocks that carry no information about the outcome, the two are not separately defined: either can absorb all of the variation in their choices that is correlated with the outcome. A comparison of the decision maker's information with the algorithm's is therefore undefined until the prediction is normalized. The independence condition of Definition 1 is the normalization used in our analysis: every factor driving their choices that is correlated with the outcome is attributed to the prediction, and everything else to the threshold. Calibration then fixes the scale of the normalized prediction.

Definition A1. A *structural model* is a joint distribution $(X, V, Y^*, \varepsilon^H, \varepsilon^T, H, T, C) \sim Q(\cdot)$ in which the decision maker observes the recorded characteristics X and private information V, the algorithm's prediction is $M = m(X)$ for a measurable $m(\cdot)$, and

i. Beliefs and loss ratios: $H = h(X, V, \varepsilon^H)$ and $T = t(X, V, \varepsilon^T)$ for measurable h taking values in $[0, 1]$ and t taking values in $(0, 1]$.

ii. Shocks: $(\varepsilon^H, \varepsilon^T) \perp\!\!\!\perp Y^* \mid (X, V)$.

iii. Threshold Rule: $C = \mathbf{1}\{H \leq T\}$ almost surely.

The decision maker's choices are *consistent with* the structural model if $Q(M = m, C = c, Y^* = y) = p_{cy}(m)$ for all $m \in \mathcal{M}$ and $c, y \in \{0, 1\}$.

The decision maker's information set is (X, V): X is what the algorithm also sees and V is what it does not. The shocks ε^H and ε^T are whatever moves their belief and their loss ratio beyond their information set: noise in the sense of Kahneman, Sibony and Sunstein (2021), heterogeneity in costs, capacity constraints, and payoffs the decision maker weighs that the researcher has not modeled (Kleinberg et al., 2017). Condition (ii) is the only restriction placed on them: given everything the decision maker knows, the shocks carry no information about the outcome. Neither h nor t is restricted, so the belief need not be accurate, and the shocks may be arbitrarily dependent on one another.

Lemma A2. *Suppose the decision maker holds the belief H and incurs the loss $L_1 = \ell_1(X, V, \varepsilon^T) \geq 0$ from selecting an individual for whom the outcome occurs, the loss $L_0 =$*

$\ell_0(X, V, \varepsilon^T) > 0$ *from passing over one for whom it does not, and no loss from a correct choice. Minimizing expected loss under their belief, with indifference resolved toward* $C = 1$*, is the threshold rule* $C = \mathbf{1}\{H \leq T\}$ *with* $T = L_0/(L_0 + L_1) \in (0, 1]$.

Proof. The expected loss under the belief is HL_1 from selecting the individual and $(1-H)L_0$ from passing them over, so the decision maker selects whenever $HL_1 \leq (1 - H)L_0$, that is, whenever $H(L_0 + L_1) \leq L_0$, which is $H \leq L_0/(L_0 + L_1)$ because $L_0 + L_1 > 0$; the ratio lies in $(0, 1]$ because $L_0 > 0$ and $L_1 \geq 0$. □

A structural model is therefore expected-loss minimization at a subjective belief, with losses that depend on the information set and on shocks that are uninformative about the outcome given it.

The decision maker's choices do not separately identify the two components of a structural model. As an example, consider a structural model $Q(\cdot)$ and move the loss ratio into the belief,

$$\tilde{H} := \frac{1 + H - T}{3}, \qquad \tilde{T} :\equiv \frac{1}{3}.$$

Since $\tilde{H} \leq 1/3$ exactly when $H \leq T$, the choices are unchanged; $\tilde{H}$ is a function of the information set and the two shocks and $\tilde{T}$ is constant, so this is again a structural model, now with a constant loss ratio. Its belief mixes what the decision maker believes with what they want, and it is calibrated only if the outcome rate given (H, T) averages to $(1+H-T)/3$ along each level set of $\tilde{H}$, a coincidence between the two. Alternatively, we can move the belief into the loss ratio instead: with $U \sim \text{Unif}(0, 1)$ drawn independently of everything,

$$\tilde{H} := \frac{1 + U}{2}, \qquad \tilde{T} := \frac{1 + 3C}{4}.$$

Since $1/2 < \tilde{H} < 1$ while $\tilde{T}$ equals 1 on $\{C = 1\}$ and $1/4$ on $\{C = 0\}$, the choices are again unchanged; the choice is a function of the information set and the shocks, so this too is a structural model, whose belief is pure noise, independent of the information set, the outcome and the choice, and never calibrated, since $Q(Y^* = 1 \mid \tilde{H})$ is constant while $\tilde{H}$ is not. Neither reassignment changes the distribution of (M, C, Y^*), so the data are consistent with the original model, with one in which every variation in the choices is belief, and with one in which none of it is. Definition 1 resolves this ambiguity by a normalization. The first reassignment has a constant threshold and so satisfies its independence condition; the second satisfies it only if $C \perp\!\!\!\perp (M, Y^*)$, because $\tilde{T}$ is a one-to-one function of the choice and $\tilde{H}$ is independent of (M, Y^*, C). The definition therefore admits the reassignment that moves variation in the loss ratio into the prediction and excludes the one that moves the belief into the threshold: the prediction is the component of the choice process that is correlated with

the outcome, which is the normalization of the footnote to Definition 1.

The structural split and the normalized one coincide when the belief is accurate given the information set, $H = Q(Y^* = 1 \mid X, V)$, and the loss ratio is independent of the information set and the outcome, $T \perp\!\!\!\perp (X, V, Y^*)$. Then (M, Y^*, H, T, C) is a threshold model: H and M are functions of (X, V), so $T \perp\!\!\!\perp (M, Y^*, H)$ and in particular $T \perp\!\!\!\perp (M, Y^*) \mid H$; calibration holds by iterated expectations; and $\rho_Q(h) = Q(T \geq h)$ is nonincreasing, left-continuous and equal to one at zero, so the model exhibits monotone choice behavior, with a fixed threshold when the loss ratio is constant. The structural models of expert decision making cited in Section 2.2 are of this kind within a coarse group of individuals. In Rambachan (2024), beliefs are accurate given the characteristics and the private information, and the threshold varies only with a researcher-specified subset of the characteristics. In Chan, Gentzkow and Yu (2022), a radiologist's belief is the Bayesian posterior probability of pneumonia given their signal, and the threshold is a radiologist-specific cost ratio. In Arnold, Dobbie and Hull (2022), a judge's belief is the Bayesian posterior probability of pretrial misconduct given a noisy signal, and the threshold varies with the judge and the defendant's race. Within the group, each is a threshold model with a fixed threshold, and the belief is the prediction of Definition 1.

The attribution problem above is the one Canay, Mogstad and Mountjoy (2024) identify for outcome tests of discrimination in generalized Roy models: a difference in thresholds across groups (taste-based discrimination), can be relabeled as a difference in beliefs across groups, so the marginal outcome test identifies taste-based discrimination only under restrictions on beliefs. Hull (2021) defines bias to include inaccurate statistical discrimination together with taste, so that everything in the decision beyond accurate beliefs counts as bias. Definition 1 is the analogue for the decision maker's prediction: everything in their choices that is correlated with the outcome counts as prediction, and the bounds of Section 4.2 measure what the data then leave open about its information frontier.

B.2 Monotone Choice Behavior and Independent Thresholds

This appendix proves the claim in the footnote to Assumption 1: a threshold that is independent of the algorithm's prediction, the outcome, and the decision maker's prediction implies monotone choice behavior, and every model with monotone choice behavior has a counterpart with an independent threshold and the same distribution of (M, Y^*, H, C). Recall the choice probability function $\rho_Q(h) = Q(C = 1 \mid H = h) = Q(T \geq h \mid H = h)$ of Section 2.3.

Lemma A3. *Let $Q(\cdot)$ be a threshold model.*

i. If $T \perp\!\!\!\perp (M, Y^*, H)$ under $Q(\cdot)$, then $Q(\cdot)$ exhibits monotone choice behavior.

ii. If $Q(\cdot)$ exhibits monotone choice behavior, there is a threshold model $\tilde{Q}(\cdot)$ with $\tilde{T} \perp\!\!\!\perp (M, Y^*, H)$ and $\tilde{Q}(\tilde{T} \geq h) = \rho_Q(h)$ for almost every h, under which (M, Y^*, H, C) has the same distribution as under $Q(\cdot)$; in particular, $\tilde{Q}(\cdot)$ is consistent with the decision maker's choices if and only if $Q(\cdot)$ is.

Proof. For (i), independence gives $\rho_Q(h) = Q(T \geq h)$, which is nonincreasing and left-continuous in h and equal to one at $h = 0$, as the footnote to Assumption 1 requires.

For (ii), let $\mathcal{R}$ be the nonincreasing, left-continuous version of ρ_Q with $\mathcal{R}(0) = 1$. These are the properties of $h \mapsto \Pr(\tilde{T} \geq h)$ for a random variable $\tilde{T}$ with values in $[0, 1]$, so such a $\tilde{T}$ exists; it equals zero with probability $1 - \lim_{h \downarrow 0} \mathcal{R}(h)$, the share of individuals never selected whatever their prediction, which is zero when the choice probability function is continuous at zero and otherwise requires the threshold to be allowed the value zero. Let $\tilde{Q}(\cdot)$ draw (M, Y^*, H) as under $Q(\cdot)$ and $\tilde{T}$ independently of it, and set $\tilde{C} := \mathbf{1}\{H \leq \tilde{T}\}$. The threshold rule holds, $\tilde{T} \perp\!\!\!\perp (M, Y^*) \mid H$ because $\tilde{T}$ is independent of (M, Y^*, H), and calibration holds because the distribution of (Y^*, H) is unchanged, so $\tilde{Q}(\cdot)$ is a threshold model. For every m and y and almost every h,

$$\begin{aligned}\tilde{Q}(\tilde{C} = 1 \mid M = m, Y^* = y, H = h) = \tilde{Q}(\tilde{T} \geq h) = \mathcal{R}(h) \\ = Q(T \geq h \mid H = h) = Q(C = 1 \mid M = m, Y^* = y, H = h),\end{aligned}$$

the last equality by the threshold rule and $T \perp\!\!\!\perp (M, Y^*) \mid H$ under $Q(\cdot)$. The two models give the same conditional distribution of the choice given (M, Y^*, H) and the same distribution of (M, Y^*, H), hence of (M, Y^*, H, C); consistency depends on the model only through the distribution of (M, C, Y^*). □

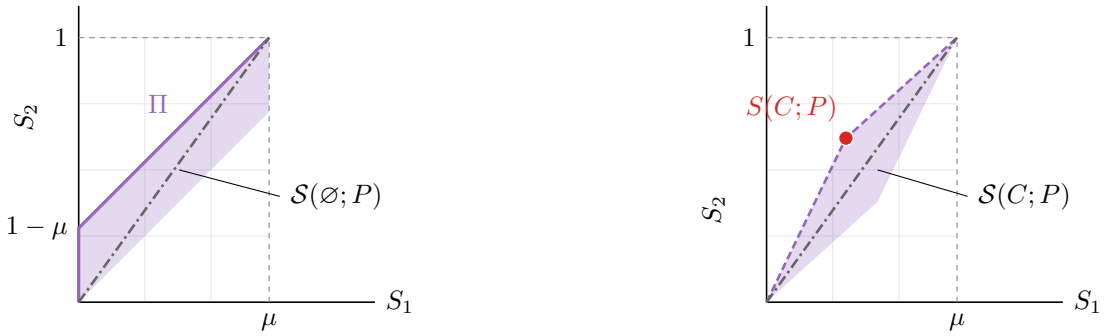
The identified set of the distribution of (M, Y^*, H, C) is therefore the same under monotone choice behavior and under an independent threshold, and so are the bounds on every object that depends on the model only through that distribution: the information frontier and the value of elicitation, but not the value of AI assistance, which depends on where the threshold lies on either side of the prediction (Section 6.1).

B.3 Properties of Scores and Values

This appendix collects the properties of feasible sets, information frontiers, and values that are referenced in Sections 4 and 5.

Fix a signal W and a threshold model $Q(\cdot)$. Write $q(W) := Q(Y^* = 1 \mid W)$ for the outcome rate given the signal, and $\mu = E_Q[Y^*]$ for the outcome rate under $Q(\cdot)$, which is $P(Y^* = 1)$ for every consistent model. Two reductions are used repeatedly. First, a decision rule d is implemented as $C^d = \mathbf{1}\{U \leq d(W)\}$ with U independent of (M, Y^*, H, T), so the

Figure A1: Two Reference Sets: Perfect Prediction and the Decision Maker's Own Choices

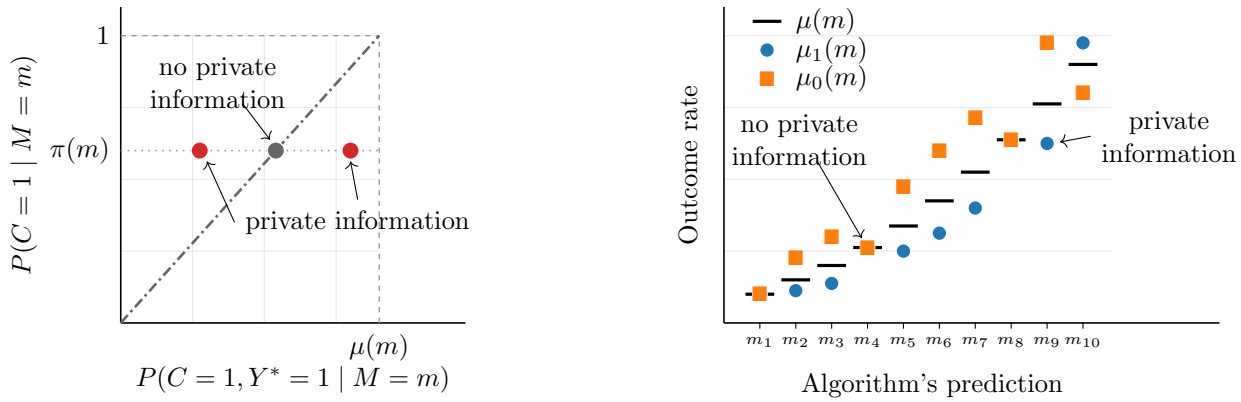


Panel A: The perfect-prediction set and the random-selection set

Panel B: The feasible set of the decision maker's own choices

Notes. Scores are $S = (S_1, S_2)$, the selected-outcome rate and the selection rate. Panel A plots the perfect-prediction set Π (shaded), the feasible set of a signal that reveals the outcome: its frontier (solid line) runs from $(0,0)$ to $(0, 1-\mu)$ and then to $(\mu, 1)$. The dash-dot segment from $(0,0)$ to $(\mu,1)$ is the random-selection set $\mathcal{S}(\varnothing; P)$, the feasible set of an uninformative signal. Panel B plots the feasible set $\mathcal{S}(C;P)$ of the decision maker's own choice (shaded), the scores of rules that follow, reverse, or randomize over their choice; its frontier (dashed line) runs from $(0,0)$ through the realized score $S(C;P)$ to $(\mu,1)$. Every feasible set contains $\mathcal{S}(\varnothing;P)$ and lies in Π. See Section 4.2 for discussion.

Figure A2: Detecting Private Information within Values of the Algorithm's Prediction

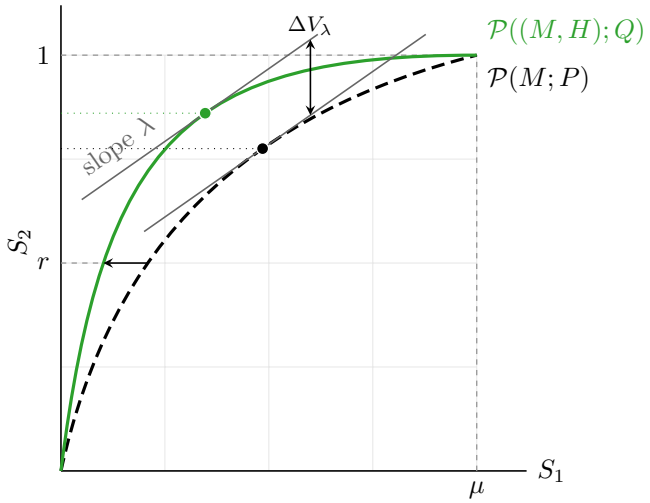


Panel A: One value of the algorithm's prediction

Panel B: Across values of the algorithm's prediction

Notes. Panel A plots, at one value m of the algorithm's prediction, the share of individuals who are selected and have the outcome against the share selected. If $C \perp\!\!\!\perp Y^* \mid M = m$, the point lies on the dash-dot segment from the origin to $(\mu(m), 1)$, the scores of selecting at random; the gray point is such a score at the realized selection rate $\pi(m)$. A point off the segment (red) shows private information at m (Lemma 2(i)). Panel B shows the same comparison across values of the algorithm's prediction: the outcome rate among the individuals the decision maker selects, $\mu_1(m)$ (circles), and among those they do not select, $\mu_0(m)$ (squares), against the outcome rate at that value, $\mu(m)$ (dashes). The decision maker has private information at the values with $\mu_1(m) \neq \mu_0(m)$. See Section 5.1 for discussion.

Figure A3: The Value of the Decision Maker's Prediction at a Weight and at a Selection Rate



Notes. Scores are $S = (S_1, S_2)$, the selected-outcome rate and the selection rate. The dashed curve is the algorithm's frontier $\mathcal{P}(M;P)$, and the solid curve the frontier $\mathcal{P}((M,H);Q)$ of the algorithm's prediction together with the decision maker's prediction under a threshold model. At a weight λ, each signal's value V_λ is the intercept of the line of slope λ tangent to its frontier, and the gain from the decision maker's prediction, $\Delta V_\lambda := V_\lambda((M,H);Q) - V_\lambda(M;P)$, is the vertical distance between the two tangents; the dotted lines mark the selection rates at the two tangency points. At a fixed selection rate r, the gain is the horizontal distance between the two frontiers (arrow). For the value of AI assistance (Section 6.1), the dashed curve is instead the frontier of the decision maker's prediction alone, $\mathcal{P}(H;Q)$, and the slope is $1/t$ at their threshold t. See Section 5.2 for discussion.

score of a rule depends on d only through $d(W)$: $Q(C^d = 1) = E_Q[d(W)]$ and $Q(C^d = 1, Y^* = 1) = E_Q[d(W)Y^*]$. Second, by iterated expectations, $E_Q[d(W)Y^*] = E_Q[d(W)q(W)]$, so

$$S(d;Q) = \big(E_Q[d(W)\,q(W)],\ E_Q[d(W)]\big) \in [0,\mu] \times [0,1]. \tag{A8}$$

The score of Section 2.4 is the pair of expected payoffs $(E_Q[C^d g_1(Y^*)], E_Q[C^d g_2(Y^*)])$ for the payoff functions $g_1(y) = y$ and $g_2(y) = 1$. The false and true positive rates, $Q(C^d = 1 \mid Y^* = 0)$ and $Q(C^d = 1 \mid Y^* = 1)$, are the pair with $g_1(y) = (1-y)/(1-\mu)$ and $g_2(y) = y/\mu$. For a general pair of payoff functions,

$$E_Q[d(W)g_k(Y^*)] = g_k(0)E_Q[d(W)] + (g_k(1) - g_k(0))E_Q[d(W)Y^*],$$

so each component of the general score is a fixed linear combination of the two components of the paper's score. Whenever the resulting linear map is invertible, as it is for the true and false positive rates, inclusion, dominance, and the information frontier (up to the orientation convention) are the same under either choice; this is the sense in which the results of Section 4 hold for the general score. We state everything for the paper's choice.

Lemma A4. *Let W be a signal and $Q(\cdot)$ a threshold model, and let $F(\cdot)$ be the law of the outcome rate $q(W)$ under $Q(\cdot)$, with quantile function $F^{-1}(\cdot)$.*

i. The feasible set $\mathcal{S}(W;Q)$ is convex and compact. It contains $(0,0)$, $(\mu,1)$, and the segment joining them, and it is symmetric about their midpoint: $s \in \mathcal{S}(W;Q)$ if and only if $(\mu,1) - s \in \mathcal{S}(W;Q)$.

ii. For $\alpha \in [0,1]$, let $\psi_F(\alpha) := \int_0^\alpha F^{-1}(u)\,du$, the integral of the outcome rate over the lowest α of its mass. Then ψ_F is convex and nondecreasing, with $\psi_F(0) = 0$ and $\psi_F(1) = \mu$, and strictly increasing on $[Q(q(W) = 0), 1]$; the feasible set is the band

$$\mathcal{S}(W;Q) = \{(s_1,\alpha) : \alpha \in [0,1],\ \psi_F(\alpha) \le s_1 \le \mu - \psi_F(1-\alpha)\},$$

and the information frontier is $\mathcal{P}(W;Q) = \{(\psi_F(\alpha),\alpha) : \alpha \in [Q(q(W) = 0), 1]\}$. A rule $d \in \mathcal{D}(W)$ has $S(d;Q) \in \mathcal{P}(W;Q)$ if and only if, for some $\tau \in [0,1]$, $d(W) = 1$ almost surely on $\{q(W) < \tau\}$ and $d(W) = 0$ almost surely on $\{q(W) > \tau\}$, and in addition $d(W) = 1$ almost surely on $\{q(W) = 0\}$.

iii. If W' is a function of W, then $\mathcal{S}(W';Q) \subseteq \mathcal{S}(W;Q)$; if the function is one-to-one, the two feasible sets coincide.

iv. W dominates W' if and only if W is more informative than W' about Y^* in the sense

of Blackwell (1953), that is, if and only if there is a Markov kernel K from the values of W to the values of W' such that $Q(W' \in \cdot \mid Y^* = y) = \int K(w, \cdot)\, Q(W \in dw \mid Y^* = y)$ for $y \in \{0, 1\}$.

v. $\mathcal{S}(W; Q) \subseteq \mathcal{S}(Y^*; Q) =: \Pi$*, the parallelogram with vertices* $(0, 0)$*,* $(0, 1-\mu)$*,* $(\mu, 1)$*, and* (μ, μ)*, which is closed and depends on the model only through* μ*.*

Proof. Proof of (i): The set $\mathcal{D}(W)$ is convex and $d \mapsto S(d; Q)$ is affine, so $\mathcal{S}(W; Q)$ is convex. It is compact because $\{d(W) : d \in \mathcal{D}(W)\}$ is weak* compact in $L^\infty(\sigma(W))$ and $Z \mapsto (E_Q[Zq(W)], E_Q[Z])$ is weak* continuous. The constant rules $d \equiv \rho$ give the scores $(\rho\mu, \rho)$, which trace the segment from $(0, 0)$ to $(\mu, 1)$. The map $d \mapsto 1 - d$ is a bijection of $\mathcal{D}(W)$ and sends $S(d; Q)$ to $(\mu, 1) - S(d; Q)$, which gives the symmetry.

Proof of (ii): Write $V := q(W)$; by equation (A8), $S(d; Q) = (E_Q[d(W)V], E_Q[d(W)])$ and $E_Q[V] = \mu$. Call d a cutoff rule in V at τ if $d(W) = 1$ almost surely on $\{V < \tau\}$ and $d(W) = 0$ almost surely on $\{V > \tau\}$. For rules d and d' with the same selection rate and any τ,

$$E_Q[d(W)V] - E_Q[d'(W)V] = E_Q\big[(d(W) - d'(W))(V - \tau)\big],$$

whose integrand is nonpositive when d is a cutoff rule in V at τ, and zero only if d' is one as well; so a cutoff rule in V minimizes the first score at its selection rate, and every minimizer at rate α is a cutoff rule in V at the quantile $\tau := F^{-1}(\alpha)$, at which a cutoff rule d^* with rate α exists. Since $F^{-1}(U)$ has law $F(\cdot)$ for U uniform on $[0, 1]$, the minimum is $E_Q[d^*(W)V] = \int_0^\alpha F^{-1}(u)\, du = \psi_F(\alpha)$. The quantile function is nonnegative and nondecreasing, so ψ_F is nondecreasing and convex with $\psi_F(0) = 0$ and $\psi_F(1) = E_Q[V] = \mu$, and strictly increasing on $[F(0), 1]$ because $F^{-1}(u) > 0$ for $u > F(0) = Q(V = 0)$. By the symmetry in (i), the largest first score at rate α is $\mu - \psi_F(1 - \alpha)$, and by convexity the feasible scores at that rate fill the interval between the two: this is the band. A feasible score above the graph of ψ_F is improved upon by the point of the graph at its rate, and a point $(0, \alpha)$ of the graph with $\alpha < F(0)$ by $(0, F(0))$; for $\alpha \geq F(0)$ an improvement (s_1', α') of $(\psi_F(\alpha), \alpha)$ would have $\psi_F(\alpha') \leq s_1' \leq \psi_F(\alpha)$ and $\alpha' \geq \alpha$, hence $\alpha' = \alpha$ by strict monotonicity and $s_1' = \psi_F(\alpha)$; so the frontier is the graph on $[F(0), 1]$. Hence $S(d; Q)$ lies on the frontier if and only if d is a cutoff rule in V at some τ with rate at least $F(0)$; a cutoff rule at $\tau > 0$ has rate at least $Q(V < \tau) \geq F(0)$ and selects $\{V = 0\} \subseteq \{V < \tau\}$, while one at $\tau = 0$ has rate $E_Q[d(W)\mathbf{1}\{V = 0\}] \leq F(0)$, with equality if and only if $d(W) = 1$ almost surely on $\{V = 0\}$.

Proof of (iii): If $W' = \phi(W)$, every rule $d' \in \mathcal{D}(W')$ yields the rule $d' \circ \phi \in \mathcal{D}(W)$ with the same score, so $\mathcal{S}(W'; Q) \subseteq \mathcal{S}(W; Q)$; if ϕ is one-to-one, $W = \phi^{-1}(W')$ and the reverse inclusion holds as well.

Proof of (iv): If W' is obtained from W through the kernel K, then for any rule $d' \in$

$\mathcal{D}(W')$ the rule $d(w) := \int d'(w')K(w, dw')$ belongs to $\mathcal{D}(W)$ and has the same score, because $E_Q[d'(W') \mid W, Y^*] = d(W)$ for both values of Y^*; hence $\mathcal{S}(W'; Q) \subseteq \mathcal{S}(W; Q)$. For the converse, the feasible set of a signal is, up to the invertible linear map of the opening paragraph, the set of pairs of error probabilities of the tests of $Y^* = 0$ against $Y^* = 1$ based on that signal, and for a binary outcome Blackwell (1953) shows that the inclusion of these sets characterizes the existence of such a kernel.

Proof of (v): For $d \in \mathcal{D}(W)$, the rule $\bar{d}(y) := E_Q[d(W) \mid Y^* = y]$ based on the outcome has the same score, by equation (A8) and iterated expectations, so $\mathcal{S}(W; Q) \subseteq \mathcal{S}(Y^*; Q)$. A rule based on the outcome selects the individuals with $Y^* = 0$ at some rate a and those with $Y^* = 1$ at some rate b, with score $(b\mu,\ b\mu + a(1-\mu))$, and as (a, b) ranges over $[0,1]^2$ these scores trace the stated parallelogram, which is closed. □

With our choice of scores, part (ii) says the information frontier is the curve traced by the cutoff rules on the outcome rate $q(W)$, running from the score of selecting exactly the individuals with no risk to the score of selecting everyone. For the decision maker's prediction, $q(H) = H$ under the threshold model, so part (ii) characterizes the rules that attain their frontier as the cutoff rules on H that select almost every individual with $H = 0$. Part (iii) is the relabeling invariance of Section 2.4, and part (iv) its Blackwell interpretation. When the signal is the algorithm's prediction, the decision maker's choice, or the pair, the joint distribution of (W, Y^*) is the same under every consistent model, so $\mathcal{S}(M; P)$, $\mathcal{S}(C; P)$, and $\mathcal{S}((M, C); P)$ are identified.

Lemma A5. *For a signal W, a threshold model $Q(\cdot)$, and a weight $\lambda > 0$, let $V_\lambda(W; Q) := \max_{d \in \mathcal{D}(W)} S_2(d; Q) - \lambda S_1(d; Q)$ be the value of the signal, as in Section 5.*

i. The maximum is attained by the rule $d^*(w) = \mathbf{1}\{q(w) < 1/\lambda\}$, and $V_\lambda(W; Q) = E_Q[(1 - \lambda\, q(W))_+]$. If $\lambda \leq 1$, then $V_\lambda(W; Q) = 1 - \lambda\mu$ for every signal.

ii. W dominates W' if and only if $V_\lambda(W; Q) \geq V_\lambda(W'; Q)$ for every $\lambda > 0$.

Proof. By equation (A8), $S_2(d; Q) - \lambda S_1(d; Q) = E_Q[d(W)(1 - \lambda q(W))]$, which is maximized pointwise by $d^*(w) = \mathbf{1}\{q(w) < 1/\lambda\}$, with value $E_Q[(1 - \lambda q(W))_+]$; if $\lambda \leq 1$ the integrand is nonnegative everywhere and the value is $1 - \lambda\mu$. For (ii), inclusion of feasible sets ranks the maxima of a common objective; conversely, $V_\lambda(W; Q)$ is the support function of $\mathcal{S}(W; Q)$ in the direction $(-\lambda, 1)$, and by the symmetry in Lemma A4(i) the support function in the direction $(\lambda, -1)$ is $\lambda\mu - 1 + V_\lambda(W; Q)$, while in the directions with both components nonnegative or both nonpositive it is the value at the corner $(\mu, 1)$ or zero, the same for every signal. Every direction is a positive multiple of one of these, so ranking the values at

every weight ranks the support functions in every direction, which for compact convex sets is inclusion. □

C Additional Details on Computing Bounds under the Threshold Model

C.1 Bounds on Predictive Skill

This appendix describes how we compute the bounds on a judge's information frontier that Section 4.3.2 reports. For a selection rate r, the bounds are the lowest release-misconduct rate that the decision maker's prediction could support under some threshold model consistent with their choices, $\underline{\tilde{S}}_H(r;P)$, and the lowest it is guaranteed to support under every such model, $\overline{\tilde{S}}_H(r;P)$: the frontier points at the rate r of the upper and lower bounds on the feasible set of their prediction in Proposition 4.2, computed under the threshold model alone, under monotone choice behavior, and under a fixed threshold. All but one of these six quantities have closed forms; the exception, the lowest rate the prediction could support under monotone choice behavior, is the optimal value of a program that we describe below.

Every quantity below is a function of the joint distribution $p_{cy}(m)$ of the algorithm's prediction, the decision maker's choice, and the outcome. In the application, the algorithm's prediction is its decile of predicted risk, so $|\mathcal{M}| = 10$, and the deciles are weighted equally, $p(m) = 1/10$. For each judge and each decile m, we estimate three quantities in Section 3.3: the judge's release rate $\pi(m)$ and release-misconduct rate $\alpha(m) := P(C = 1, Y^* = 1 \mid M = m)$ among the cases in the decile, each strata-adjusted at the decile's own mix of strata, and the decile's outcome rate $\mu(m)$, extrapolated across judges. Since $\mu(m) = \alpha(m) + P(C = 0, Y^* = 1 \mid M = m)$, the four masses at decile m are

$$p_{11}(m) = \tfrac{1}{10}\,\alpha(m), \qquad p_{10}(m) = \tfrac{1}{10}\big(\pi(m) - \alpha(m)\big),$$
$$p_{01}(m) = \tfrac{1}{10}\big(\mu(m) - \alpha(m)\big), \qquad p_{00}(m) = \tfrac{1}{10}\big(1 - \pi(m) - \mu(m) + \alpha(m)\big),$$

and the selection rate at which we make the comparison is the judge's own release rate, $r = \frac{1}{10}\sum_m \pi(m)$.

To calculate the bounds, we consider the optimal values of programs whose unknowns describe a threshold model on finitely many values of the decision maker's prediction. We allow for K distinct values $0 \le h_1 < \cdots < h_K \le 1$ of their prediction. We define the mass $w_{my,k} \ge 0$ of individuals at the algorithm's prediction $m \in \mathcal{M}$ with outcome $y \in \{0, 1\}$ whose prediction takes the value h_k; and the probability $\rho_k \in [0, 1]$ that the decision maker

selects an individual whose prediction is h_k. The unknowns must match the data,

$$\sum_{k=1}^{K} w_{my,k} = p_{1y}(m) + p_{0y}(m), \qquad \sum_{k=1}^{K} \rho_k \, w_{my,k} = p_{1y}(m) \qquad \text{for all } m \in \mathcal{M} \text{ and } y \in \{0,1\}, \tag{A9}$$

and satisfy calibration and the threshold rule's condition at zero,

$$(1 - h_k) \sum_{m \in \mathcal{M}} w_{m1,k} = h_k \sum_{m \in \mathcal{M}} w_{m0,k} \quad \text{for all } k \leq K, \qquad \text{and} \qquad \rho_k = 1 \ \text{ if } h_k = 0. \tag{A10}$$

Under monotone choice behavior, the choice probabilities satisfy $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_K$; under a fixed threshold, $\rho_k = \mathbf{1}\{h_k \leq t\}$ for a single $t \in (0,1]$. A value carrying no mass is unconstrained apart from these orderings. Writing $w_{m,k} := w_{m1,k} + w_{m0,k}$ for the mass at prediction m and value k, the value of the decision maker's prediction at a weight $\lambda > 1$ is

$$v_\lambda^H(w) := \sum_{k=1}^{K} \Big(\sum_{m \in \mathcal{M}} \big(w_{m,k} - \lambda \, w_{m1,k} \big) \Big)_+, \tag{A11}$$

and its frontier point at a selection rate r is

$$\psi^H(r; w) := \min \Big\{ \sum_{k=1}^{K} d_k \sum_{m \in \mathcal{M}} w_{m1,k} \ : \ d_k \in [0,1], \ \sum_{k=1}^{K} d_k \sum_{m \in \mathcal{M}} w_{m,k} = r \Big\}. \tag{A12}$$

A point satisfying these constraints describes a threshold model consistent with the decision maker's choices. The first constraints in equation (A9) match the masses $P(M = m, Y^* = y)$ and the second match the masses $p_{1y}(m)$ of the selected individuals. The calibration constraint in equation (A10) says that the outcome occurs at rate h_k among the individuals whose prediction is h_k; and the condition at zero is the threshold rule's requirement that every individual with prediction zero is selected. The model's choice probability function takes the value ρ_k at h_k, so the ordering constraint is monotone choice behavior and the step constraint is a fixed threshold at t. The objective at a weight is the value of the decision maker's prediction under this model (Lemma A5), with the positive part taken of the sum across the algorithm's predictions at each value, because a rule based on the prediction alone cannot separate the algorithm's predictions at the same value; and the objective at a rate is the prediction's frontier point, the inner program choosing the rule d_k over the values.

Computation: Write $\bar{\pi} := P(C = 1)$ for the judge's release rate. The lowest release-misconduct rate the prediction is guaranteed to support is the frontier point of the judge's

own choices, the lower bound $\mathcal{S}(C;P)$ of Proposition 4.2, under every restriction. When $\mu_1 \leq \mu_0$, the rule that selects the individuals the judge selected before those they did not attains it,

$$\overline{\tilde{S}}_H(r;P) = \mu_1 \min\{r, \bar{\pi}\} + \mu_0 \max\{0,\, r - \bar{\pi}\},$$

and under the threshold model alone with $\mu_0 < \mu_1$ the rule that selects the individuals the judge did not select first attains it, $\mu_0 \min\{r, 1-\bar{\pi}\} + \mu_1 \max\{0,\, r-(1-\bar{\pi})\}$. In the program, the point with two values, μ_1 carrying the masses $p_{1y}(m)$ with $\rho = 1$ and μ_0 carrying the masses $p_{0y}(m)$ with $\rho = 0$, attains it, and satisfies monotone choice behavior and a fixed threshold when $\mu_1 < \mu_0$, as in Appendix C.2. The lowest rate the prediction could support is, under the threshold model alone, the perfect predictor's frontier point $\max\{0,\, r-(1-\mu)\}$, the upper bound Π of Proposition 4.2; and under a fixed threshold, which requires $\mu_1 < \mu_0$,

$$\underline{\tilde{S}}_H^{\mathrm{FT}}(r;P) = \begin{cases} \max\{0,\ \bar{\pi}\mu_1 - (\bar{\pi} - r)\,\mu_0\} & \text{for } r \leq \bar{\pi}, \\ \max\{\mu_1 r,\ r - (1-\mu)\} & \text{for } r \geq \bar{\pi}, \end{cases}$$

the frontier of the union of the feasible sets of two calibrated predictions that each take two values, one with the values 0 and μ_0 and one with the values μ_1 and 1 (Appendix A.2). The one quantity without a closed form is the lowest rate the prediction could support under monotone choice behavior, which we compute next.

Under monotone choice behavior, the lowest rate the prediction could support at a selection rate r is attained among models whose prediction takes at most four values, so the program reduces to one with four values, one choice probability, and $8|\mathcal{M}|$ masses.[13] Write $W_k := \sum_{m\in\mathcal{M}} \sum_y w_{my,k}$ and $A_k := \sum_{m\in\mathcal{M}} w_{m1,k}$ for the mass and the outcome mass at the value $k \in \{a, b, c, d\}$. The four values carry the choice probabilities 1, ρ, ρ, and 0, and the

[13] Take a consistent model and a rule attaining its frontier point at r, with cutoff τ in the prediction and ρ the judge's choice probability at τ. Split the individuals below τ into two groups that the judge selects with probabilities one and ρ, and those above into two groups with probabilities ρ and zero, and replace each group's predictions by its outcome rate; the data, the frontier point, calibration, and monotone choice behavior are preserved, and the prediction now takes four values. We omit the formal statement.

rule selects the first two, so

$$\begin{aligned}
\underline{\tilde{S}}_H^{\mathrm{MC}}(r;P) = \min_{w,\rho,h} \quad & A_a + A_b \\
\text{subject to} \quad & \sum_k w_{my,k} = p_{1y}(m) + p_{0y}(m) \qquad \text{for all } m \text{ and } y, \\
& w_{my,a} + \rho\left(w_{my,b} + w_{my,c}\right) = p_{1y}(m) \qquad \text{for all } m \text{ and } y, \\
& W_a + W_b = r, \qquad h_k W_k = A_k \quad \text{for } k \in \{a,b,c,d\}, \\
& h_a \le h_b \le h_c \le h_d, \\
& w_{my,k} \ge 0, \qquad \rho \in [0,1], \qquad h_k \in [0,1].
\end{aligned} \tag{A13}$$

The first two constraints match the data, the third fixes the selection rate, the fourth makes each value the outcome rate of the mass placed at it, and the orderings are monotone choice behavior. Its only nonconvex unknowns are ρ and the four values. The program with $K = 41$ free values could not be solved to certified optimality: the solver's relaxation returns the perfect predictor's frontier point, so its best feasible point cannot be told from a suboptimal one. The four-value program has bilinear terms in five scalars only, and a solver for nonconvex quadratic programs certifies it in about a second per judge; the perfect predictor's frontier point is a floor on every value we report.

Empirical application: We solve the four-value program judge by judge, with the judge's own masses as inputs, for the 362judges with cases in every decile of predicted risk; it applies to the 223judges whose choices are consistent with monotone choice behavior, those with $\mu_1 < \mu_0$, and the other five bounds come from the closed forms. Figure 4 and Table 2 report the bounds at each judge's own release rate, and Figure 5 traces them for Judges A, B, and C at selection rates in steps of 0.01. The brackets in Table 2 are 90 percent credible intervals over 500 Bayesian-bootstrap draws of the judge's masses, with the closed forms recomputed and the four-value program re-solved at each draw's own release rate.

C.2 Computing the Bounds on the Value of Elicitation

This appendix describes how we compute the bounds on the value of eliciting the decision maker's prediction reported in Section 5.3.2. That section compares three release-misconduct rates at the judge's release rate r, which in the notation of Section 4.2 are the frontier points $\tilde{S}_M(r;P)$, $\tilde{S}_{(M,C)}(r;P)$, and $\tilde{S}_{(M,H)}(r;Q)$. The first two are identified. The third depends on the threshold model, and we report the lowest release-misconduct rate the data cannot rule out, $\underline{\tilde{S}}_{(M,H)}(r;P)$, under the threshold model alone, under monotone choice behavior, and under a fixed threshold; it is the fixed-rate form of the upper bound on the value of elicitation in Proposition 5.1. The highest rate, the fixed-rate form of the lower bound on

the value, is $\tilde{S}_{(M,C)}(r;P)$ under every restriction, as we show below.

The values of the two identified signals at a weight $\lambda > 1$ are

$$V_\lambda(M;P) = \sum_{m\in\mathcal{M}} p(m)\big(1-\lambda\,\mu(m)\big)_+,$$

and

$$V_\lambda\big((M,C);P\big) = \sum_{m\in\mathcal{M}} p(m)\Big[\pi(m)\big(1-\lambda\,\mu_1(m)\big)_+ + \big(1-\pi(m)\big)\big(1-\lambda\,\mu_0(m)\big)_+\Big].$$

At a selection rate, the best rule based on a signal selects in increasing order of the outcome probability given the signal, taking the fraction of the last group needed to reach the rate (Lemma A4(ii)). Writing $p_c(m) := p_{c1}(m) + p_{c0}(m)$ for the mass of individuals at prediction m with choice c,

$$\tilde{S}_{(M,C)}(r;P) = \min\Big\{\sum_{m\in\mathcal{M}}\sum_{c\in\{0,1\}} d_{mc}\,p_{c1}(m)\ :\ d_{mc}\in[0,1],\ \sum_{m\in\mathcal{M}}\sum_{c\in\{0,1\}} d_{mc}\,p_c(m) = r\Big\},$$

a linear program solved by selecting the groups of selected and of unselected individuals at each prediction in increasing order of $\mu_1(m)$ and $\mu_0(m)$, and $\tilde{S}_M(r;P)$ is the same program with the rule indexed by the prediction alone, solved by selecting the predictions in increasing order of $\mu(m)$.

The third frontier point, $\tilde{S}_{(M,H)}(r;Q)$, depends on the threshold model. Its bounds are the optimal values of the program with the unknowns and constraints of Appendix C.1, equations (A9) and (A10), and the objectives of the pair. The designer holds the pair and selects prediction by prediction, so the value of the pair at a weight $\lambda > 1$ is

$$v_\lambda^{(M,H)}(w) := \sum_{m\in\mathcal{M}}\sum_{k=1}^{K}\big(w_{m,k} - \lambda\,w_{m1,k}\big)_+, \tag{A14}$$

and its frontier point at a selection rate r is

$$\psi^{(M,H)}(r;w) := \min\Big\{\sum_{m\in\mathcal{M}}\sum_{k=1}^{K} d_{mk}\,w_{m1,k}\ :\ d_{mk}\in[0,1],\ \sum_{m\in\mathcal{M}}\sum_{k=1}^{K} d_{mk}\,w_{m,k} = r\Big\}, \tag{A15}$$

the inner program choosing the designer's rule d_{mk} over the pairs of a prediction and a value (Lemma A5). The bounds of Section 5.2 are the infimum and the supremum of these objectives over the constraint set.

Computation: Two of the three bounds in each column of Table 4 have closed forms, so the programs are solved only for the upper bound on the value, the lowest release-misconduct rate the pair could attain, under monotone choice behavior and under a fixed threshold. The lower bound on the value has a closed form under every restriction: the highest release-misconduct rate the pair could have is that of the algorithm's prediction with the choice, $\tilde{S}_{(M,C)}(r; P)$. For any point satisfying the constraints, the rule $d_{mk} := \rho_k\, d_{m1} + (1 - \rho_k)\, d_{m0}$ reproduces the score of any rule (d_{m1}, d_{m0}) based on the algorithm's prediction and the choice, because the choice constraints in equation (A9) give

$$\sum_{k=1}^{K} d_{mk}\, w_{my,k} = d_{m1}\, p_{1y}(m) + d_{m0}\, p_{0y}(m) \qquad \text{for all } m \in \mathcal{M} \text{ and } y \in \{0, 1\},$$

so $\psi^{(M,H)}(r; w) \leq \tilde{S}_{(M,C)}(r; P)$ at every feasible point: the pair never does worse than the choice. The point with two values attains the bound: $h_1 = \mu_1$ carries the masses $p_{1y}(m)$ with $\rho_1 = 1$, and $h_2 = \mu_0$ carries the masses $p_{0y}(m)$ with $\rho_2 = 0$, where μ_1 and μ_0 are the outcome rates among the individuals the decision maker selects and does not select. It satisfies the constraints whenever $\mu_1 \neq \mu_0$, since the outcome rate at each value is the value itself, it satisfies monotone choice behavior and a fixed threshold whenever $\mu_1 < \mu_0$, which Corollary 2.1 requires for those classes to be nonempty, and its pair is a relabeling of (M, C). The lower bound on the value is therefore the value of the algorithm's prediction with the choice under all three restrictions, as in Proposition 5.1, attained whenever $\mu_1 \neq \mu_0$ and approached otherwise.

The upper bound on the value has a closed form under the threshold model alone: the lowest release-misconduct rate the pair could attain is the perfect predictor's. Every rule with selection rate r has first score $E_Q[dY^*] \geq E_Q[d] - E_Q[1 - Y^*] = r - (1 - \mu)$ and at least zero, so $\psi^{(M,H)}(r; w) \geq \max\{0,\, r - (1 - \mu)\}$, the frontier point of the perfect-prediction set at the rate r. Under the threshold model alone this bound is approached: the models of Proposition 5.1 bring the pair's value arbitrarily close to the value of perfect prediction, hence its scores to the corner $(0, 1 - \mu)$ of the perfect-prediction set, and by the convexity and symmetry of the feasible set (Lemma A4(i)) its frontier point at every rate to the perfect predictor's. Under a fixed threshold the same holds whenever $\underline{T} \leq \overline{T}$ (Proposition 5.1(ii) and Lemma 3), and under monotone choice behavior whenever $\underline{T} \leq \overline{T}$ or there is no reversal across the algorithm's predictions (Proposition 5.1(i)). Otherwise the frontier point lies strictly above the perfect predictor's at some selection rates, though not necessarily at a given one, and the program must be solved.

Under monotone choice behavior we solve the program as written, with $K = 41$ values

whose locations are unknowns, minimizing jointly over the values, the masses, the choice probabilities, and the designer's rule, since the upper bound on the value is attained at the lowest release-misconduct rate, the infimum of the minimum in equation (A15). Three families of products then appear: the choice probabilities with the masses in the choice constraints, the values with the masses in the calibration constraint, and the rule with the masses in the objective and in the rate constraint. We solve the program to certified global optimality with a solver for nonconvex quadratic programs.

Under a fixed threshold the choice probabilities are a step in the ordered values, and we solve the program by enumerating the position of the step. For each $k \in \{0, 1, \ldots, K\}$, set $\rho_j = \mathbf{1}\{j \leq k\}$, so that the choice constraints become linear in the masses, $\sum_{j \leq k} w_{my,j} = p_{1y}(m)$ for all m and y, and the calibration constraint and the objective are the only bilinear parts; the lowest rate, and so the upper bound on the value, is the smallest of the $K + 1$ program values. The threshold itself is not an unknown: any t between the k-th value and the next gives the same step.

Empirical illustration: We solve the programs judge by judge, with the judge's own masses $p_{cy}(m)$ as inputs and the judge's own release rate as r, for the 362judges with cases in every decile of predicted risk; the two restrictions apply to the 223judges whose choices are consistent with monotone choice behavior. Figure 8 and Table 4 report the three release-misconduct rates at each judge's release rate, and Figure 9 repeats the lowest rate the pair could attain at selection rates in steps of 0.02 for Judges A and B; Judge C's choices are inconsistent with both restrictions. Every solve under either restriction returned the perfect predictor's frontier point, so the upper bound on the value of elicitation is the value of perfect prediction for every judge and at every selection rate we computed. This is what Proposition 5.1 predicts when $\underline{T} \leq \overline{T}$, a condition that can be checked from the data without solving a program, and the result is certified whatever the solver's optimality gap, because the perfect predictor's frontier point is a lower bound on every program value. In the application the programs therefore confirmed the closed forms above, and the only entries of Table 4 that required computation are the identified rates of the algorithm and of the algorithm with the judge's choice.

C.3 Computing the Bounds on the Value of AI Assistance

This appendix describes how we compute the bounds on the value of AI assistance reported in Section 6.2. It reports the value in two ways: at the judge's release rate, $\tilde{S}_H(r; Q) - \tilde{S}_{(M,H)}(r; Q)$ under one model, and at the judge's own thresholds, $\Delta(Q)$ in release-misconduct units, where at a threshold t a signal W is worth $t\, V_{1/t}(W; Q) = E_Q[(t - Q(Y^* = 1 \mid W))_+]$. The first measure depends on a model only through the joint distribution of the

algorithm's prediction, the outcome, and the judge's prediction, so its bounds are computed with the unknowns and constraints of Appendix C.1 and the two frontier points $\psi^{H}(r;w)$ and $\psi^{(M,H)}(r;w)$ of equations (A12) and (A15), evaluated at one feasible point. The second depends in addition on the threshold's distribution at each value of the prediction: at least h_k among the individuals the judge selects there, and below h_k among those they do not. Write θ_k for the distribution of the threshold at the k-th value, a mixture that places weight ρ_k on a distribution over $[h_k, 1]$ and weight $1-\rho_k$ on a distribution over $(0, h_k)$, and $v^{H}_{1/t}(w)$ and $v^{(M,H)}_{1/t}(w)$ for the values of the prediction and of the pair at the weight $1/t$ in equations (A11) and (A14). The value of AI assistance at a feasible point is

$$\delta(w,\theta) := \sum_{k=1}^{K}\Big(\sum_{m\in\mathcal{M}} w_{m,k}\Big)\int t\Big(v^{(M,H)}_{1/t}(w) - v^{H}_{1/t}(w)\Big)\,\theta_k(dt), \tag{A16}$$

the gain at each threshold the judge uses, in release-misconduct units, weighted by the probability that they use it, and its bounds are the infimum and the supremum of δ over the feasible points and the admissible threshold distributions. Under a fixed threshold, the threshold is a single unknown $t \in [\mu_1, \mu_0)$, common to every value, no distribution enters, and $\delta = t\big(v^{(M,H)}_{1/t}(w) - v^{H}_{1/t}(w)\big)$.

At the judge's release rate, the bounds have closed forms except under monotone choice behavior. Write $S := \bar{\pi}\mu_1$ for the judge's realized release-misconduct rate, $c := \tilde{S}_{(M,C)}(\bar{\pi};P)$ for the rate of the algorithm's prediction with the judge's choice, and $o := \max\{0, \bar{\pi}-(1-\mu)\}$ for the perfect predictor's, all at the judge's release rate $\bar{\pi}$. Under every consistent model the pair does at least as well as the prediction alone and as the algorithm's prediction with the choice, and the prediction does at least as well as the choice, so

$$o \le \tilde{S}_{(M,H)}(\bar{\pi};Q) \le \min\big\{\tilde{S}_H(\bar{\pi};Q),\, c\big\}, \qquad \tilde{S}_H(\bar{\pi};Q) \le S,$$

and the value of AI assistance at the judge's rate lies in $[0, S-o]$. Under the threshold model alone, the lower bound is zero and the upper bound is $S-o$ whenever $\mu_1 < \mu_0$ and $\underline{T} \le \overline{T}$, a condition that is sufficient and not necessary at a fixed rate. Under a fixed threshold, the prediction's frontier point at the judge's rate is the realized rate S in every model (Proposition 4.1), so the lower bound is $S-c$ and the upper bound is S minus the lowest rate the pair could attain, which Appendix C.2 computes by enumerating the step position. Under monotone choice behavior, the upper bound is $S-o$ under the same condition, and the lower bound is the optimal value of the program we describe next.

The four-group reduction of Appendix C.1 carries over. It replaces a model by one whose prediction takes four values, keeping the data and the prediction's frontier point; the groups

are drawn from the prediction alone, so the pair of the reduced model is a garbling of the original pair, and its frontier point is weakly higher. The reduction therefore weakly lowers the value of AI assistance at the judge's rate, and four-value models suffice for its lower bound under monotone choice behavior and for the largest rate of the pair consistent with a given rate of the prediction alone. At a four-value point, the pair's frontier point is the inner linear program over the rule d_{mk} indexed by the algorithm's prediction and the group; replacing it by its dual, with a multiplier $\nu \in [0,1]$ on the rate constraint and $\eta_{mk} \geq 0$ on the caps $d_{mk} \leq 1$, turns the minimum over models of that minimum into a single minimization,

$$\begin{aligned} \min_{w,\rho,h,\nu,\eta} \quad & A_a + A_b - \Big(\nu\bar{\pi} - \sum_{m\in\mathcal{M}} \sum_{k} \eta_{mk}\Big) \\ \text{subject to} \quad & \text{the constraints of equation (A13) at } r = \bar{\pi}, \\ & \nu\, w_{m,k} - \eta_{mk} \leq w_{m1,k}, \qquad \eta_{mk} \geq 0, \qquad 0 \leq \nu \leq 1 \qquad \text{for all } m \text{ and } k, \end{aligned} \tag{A17}$$

where ν is the outcome probability at which the pair's rule stops selecting, and strong duality for the inner program makes the dual value equal to the pair's frontier point at every feasible point. Fixing the prediction's frontier point at a value v, by adding $A_a + A_b = v$, and maximizing the dual value in place of the objective gives the largest rate of the pair consistent with v, which the reduction makes exact; the smallest rate of the pair at a given v is found by a search over four-value models with $A_a + A_b$ within a small tolerance of v, which the reduction does not make exact, and in the application the search reached the perfect predictor's point. These two boundaries are the curves of Figure 11. The upper bound on the value is the endpoint of the second at $v = S$: when the search finds a model with monotone choice behavior whose prediction's frontier point is the realized rate and whose pair's is the perfect predictor's, the upper bound is $S - o$ under monotone choice behavior and, since the model is also consistent with the threshold model alone, under it as well.

At the judge's own thresholds the gain at a threshold t, in release-misconduct units, is $g(t) := t\, V_{1/t}\big((M,H);Q\big) - t\, V_{1/t}(H;Q)$, and the value of AI assistance is its average over the judge's threshold, the objective $\delta(w,\theta)$ of equation (A16). The lower bound is zero under the threshold model alone and under monotone choice behavior (Proposition 6.1). Under a fixed threshold $t \in [\mu_1, \mu_0)$, the prediction is worth $t\bar{\pi} - S$ in every model, the value of the judge's own choice, so the lower bound is

$$\min_{t\in[\mu_1,\mu_0]} \Big[\sum_{m\in\mathcal{M}} \sum_{c\in\{0,1\}} \big(t\, p_c(m) - p_{c1}(m)\big)_+ - \big(t\bar{\pi} - S\big) \Big], \tag{A18}$$

the gain from the algorithm's prediction with the choice at the least favorable threshold,

attained by the two-value model; the function is piecewise linear in t, so the minimum is found at the endpoints of the interval and at the outcome rates $\mu_c(m)$ inside it. For the upper bound, at any threshold the pair is worth at most $t(1-\mu)$, the value of perfect prediction, and the prediction at least what the choice is worth, so

$$g(t) \leq \bar{g}(t) := t(1-\mu) - \big(t\bar{\pi} - S\big)_+ - \big(t(1-\bar{\pi}) - (\mu - S)\big)_+,$$

a piecewise linear function with kinks at μ_1 and μ_0, whose maximum B over $[0,1]$ is attained at one of 0, μ_1, μ_0, and 1; every upper bound is at most B. Under a fixed threshold, $\bar{g}$ is linear on $[\mu_1, \mu_0]$, where it equals $S+t(1-\mu-\bar{\pi})$, and its maximum there is $\max\{\mu_1(1-\mu),\ \mu(1-\mu_0)\}$, which is B again. The cap is approached by models with a fixed threshold just inside the maximizing endpoint of $[\mu_1, \mu_0]$ whenever the thresholds just inside that endpoint are revealing (Lemma 3), which at μ_1 is the condition of Proposition 6.1; we report B as the upper bound under every restriction. For a judge outside that condition, we search for the largest value directly: the prediction takes $K = 8$ values among the individuals the judge selects and 8 among those they do not, each value's threshold is placed where the gain is largest on the side the choice determines, and the objective of equation (A16) with a point mass at each value is maximized subject to the data. The value found is a lower bound on the upper bound, and Figure 10 reports it as the best value found numerically.

Empirical illustration: We compute both measures judge by judge for the 223judges consistent with a fixed threshold, with the judge's own masses and release rate as inputs: at the judge's rate, from the closed forms above and, under monotone choice behavior, from the program of equation (A17); at the judge's thresholds, from equation (A18), the cap B, and, for Judge C, the search. Figure 10 reports both measures for Judges A, B, and C under each restriction consistent with the judge's choices, and Figure 11 traces the two boundaries of the previous paragraph for Judges A and B under monotone choice behavior across the rates the prediction alone could achieve. Table 5 reports means across judges with 90 percent credible intervals over 500 Bayesian-bootstrap draws of the judge's masses, with the closed forms recomputed and the programs re-solved at each draw's own release rate, except that the lower bound under monotone choice behavior at the judge's rate is a point estimate, the program of equation (A17) not having been re-solved on the draws.

D Sample Construction

This appendix describes the data and the construction of the estimation sample used throughout the paper (Section 3.1). Section D.1 describes the data. Section D.2 describes each sample restriction and the reason for it, Section D.3 defines the variables, Section D.4 defines

the subsamples of judges used in Sections 4 to 6, and Section D.5 compares the subsamples with the estimation sample.

D.1 Data

We use administrative records from the New York State Office of Court Administration covering the universe of arraignments in the state's criminal courts from 2020 to 2025. The unit of observation is a case at its first arraignment. Each record contains the court, the date of arraignment, the judge who conducted the arraignment, the release decision, the top arraignment charge, the defendant's demographics, prior convictions and pending cases, and the dates of any failure to appear or rearrest before the case is resolved. Cases are identified by an internal case identifier. Cases that appear twice in the records because they were arraigned twice (1 in the estimation sample) keep their earliest arraignment (the last row of Table D.1).

D.2 Sample Restrictions

Table D.1 reports the number of cases, released cases and released cases with pretrial misconduct remaining after each restriction, in the order in which we apply them. The restrictions fall into two groups. Those in Panel A remove cases in which no judge made a release decision that we can model. Those in Panel B are choices about the scope of the analysis and the requirements of the empirical strategy.

Cases without a release decision by a judge: We first drop misdemeanor cases that are not bail eligible. Under the bail reform that took effect in January 2020, a judge must release a defendant charged with such an offense, so there is no decision to study. A misdemeanor case is bail eligible if the defendant has a pending case or the top charge is a sex offense. This is the largest restriction: it removes 786,953 of the 1,585,624 recorded arraignments.

We next drop cases handled by the clerk's office, by visiting judges or hearing examiners, and cases with no assigned judge, because the assigned judge is the source of variation in Section 3.3 and these records do not identify one. We drop cases in superior courts, which hear cases after indictment rather than at first arraignment, and cases disposed of at arraignment, which have no pretrial period. We drop the 2,968 cases whose release decision is unknown. Finally, we drop the 36,476 cases with a cash bail of one dollar. A one-dollar bail is attached to a case when the defendant is already held on another case, so the recorded decision is not a release decision about the case itself. These restrictions leave 523,396 cases.

Analysis sample: We keep Black and white defendants, who account for 89.9 percent of the remaining cases, and cases whose top charge is a felony or a misdemeanor, which removes

violations and infractions. We then impose two requirements of the empirical strategy. Each court-by-month cell must contain at least 100 cases and at least 2 judges, so that judges can be compared within cells, and each judge must decide at least 100 cases. The cell requirement removes 82,981 cases and most of the judges in the smaller courts; after both requirements the sample has 382,635 cases decided by 371 judges in 13 courts. These analysis restrictions follow Rambachan (2024), who applies the same thresholds to arraignments in New York City (see also Kleinberg et al., 2017); that paper's cells are courtroom by shift, day of week, month and year, while ours are court by month because the statewide records do not identify courtrooms or shifts.

Table D.1: Sample Construction

Cut	Obs.	Share	Released	Share	Misconduct	Share
Full sample	1,585,624	1.000	1,162,245	1.000	328,098	1.000
Panel A: Structurally required						
Drop bail-ineligible misdemeanors	798,671	0.504	520,672	0.448	223,658	0.682
Drop cases handled by clerk's office	782,645	0.494	511,865	0.440	220,050	0.671
Drop cases handled by visiting judges	773,646	0.488	506,547	0.436	217,408	0.663
Drop cases with no assigned judge	773,378	0.488	506,417	0.436	217,343	0.662
Drop superior courts	609,234	0.384	429,235	0.369	193,911	0.591
Eliminate cases disposed at arraignment	562,840	0.355	414,638	0.357	193,911	0.591
Drop cases with an unknown release decision	559,872	0.353	414,567	0.357	193,156	0.589
Drop cases with a $1 cash bail	523,396	0.330	411,046	0.354	183,950	0.561
Panel B: Analysis decisions						
Restrict to black and white defendants	470,769	0.297	367,084	0.316	167,070	0.509
Restrict to felonies and misdemeanors	470,278	0.297	366,647	0.315	166,890	0.509
Restrict to court-time cells with at least 100 cases and 2 judges	387,297	0.244	308,169	0.265	137,969	0.421
Restrict to judges with at least 100 cases	382,636	0.241	304,479	0.262	136,335	0.416
Drop duplicate arraignments of the same case	382,635	0.241	304,478	0.262	136,335	0.416

Notes. Each row reports the cases that remain after the restriction in that row and every row above it. Obs. is the number of arraignments, Released the number of released defendants, and Misconduct the number of cases with pretrial misconduct; each Share column divides the count by its value in the full sample. Panel A drops cases without a release decision that our framework can study, and Panel B sets the scope of the analysis. The last row is the estimation sample. Section D.2 discusses each restriction.

D.3 Variables

Release: A defendant is released ($C = 1$) if released before trial: on recognizance, under nonmonetary conditions, or after posting bail. Defendants who had bail set and did not post it, and defendants remanded to custody, are detained. Under this definition, 79.6 percent of defendants in the estimation sample are released.

Pretrial misconduct: The main outcome is a failure to appear at a required court date or a rearrest before the case is resolved. Appendix I uses two alternatives: a failure to appear alone, and a failure to appear or a rearrest for a violent felony offense. All three are

Table D.2: Defendant Characteristics

	Full Sample (1)	Released (2)	Pretrial Misconduct (3)
Total Cases	382,635	304,478	124,352
Panel A: Top Arraignment Charge			
Misdemeanor	0.384	0.444	0.564
Felony	0.616	0.556	0.436
Violent felony	0.213	0.176	0.119
Panel B: Criminal History			
Any prior conviction	0.570	0.532	0.641
Any pending case	0.620	0.618	0.795
Panel C: Demographics			
Age	35.2	35.2	35.8
Female	0.156	0.174	0.160
White	0.385	0.398	0.404
Black	0.615	0.602	0.596

Notes. Means of defendant characteristics in the estimation sample for all defendants (column 1), released defendants (column 2), and released defendants who commit pretrial misconduct (column 3). The first row counts cases. Age is in years; every other row is a share of cases. Violent felonies are a subset of felonies. Prior convictions and pending cases are defined in Section D.3, and pretrial misconduct is a failure to appear or a rearrest before the case is resolved. Section 3.1 discusses the table.

observed only for released defendants. Among released defendants in the estimation sample, 40.8 percent commit misconduct.

Court-by-month cells: A cell is a court and the month of first arraignment. The estimation sample has 723 cells.

Characteristics: The characteristics used by the algorithm are listed in Table E.1.

D.4 Subsamples of Judges

Different parts of the analysis use different subsamples of judges, which Table D.3 collects. The comparisons in Sections 4.3.1 and 5.3.1 use all 371 judges. The bounds of Sections 4.3.2, 5.3.2 and 6.2 are computed within each decile of the algorithm's score, so they require a judge to decide cases in all ten deciles; 362 judges do so and decide at least 50 cases. Of these, 223 satisfy $\mu_1 < \mu_0$ and are therefore consistent with calibration and monotone choices (Corollary 2.1).

Table D.3: Judges and Cases in Each Analysis

	Judges	Cases
Estimation sample	371	382,635
Judges who hear cases in all ten deciles of the score	362	375,801
Of these, judges consistent with calibration and monotone choices	223	245,394

Notes. Deciles are those of the algorithm's score in the estimation sample. Cases are the cases decided by the judges in each row. Section D.4 describes where each subsample is used.

D.5 Comparison of Subsamples

Table D.4 compares the cases of the judges in each subsample with the estimation sample, and splits the estimation sample by race. The 362 judges who hear all ten deciles release 80.0 percent of their cases, against 79.6 percent in the estimation sample, and the misconduct rate among their released defendants is 41.0 percent, against 40.8 percent. The 223 judges consistent with monotone choices release 79.6 percent, with a misconduct rate of 39.7 percent among released defendants. Black defendants are released at a rate of 77.8 percent and white defendants at 82.4 percent; 33.8 and 27.9 percent face a violent top charge, 60.3 and 51.6 percent have a prior conviction, and when released 40.5 and 41.4 percent commit misconduct.

Table D.4: Characteristics of the Estimation Sample and the Analysis Subsamples

	Estimation sample	Judges who hear all ten deciles	Judges consistent with monotone choices	Black defendants	White defendants
	(1)	(2)	(3)	(4)	(5)
Released	0.796	0.800	0.796	0.778	0.824
Misconduct among released	0.408	0.410	0.397	0.405	0.414
Felony top charge	0.616	0.614	0.621	0.628	0.597
Violent top charge	0.315	0.317	0.316	0.338	0.279
Any prior conviction	0.570	0.567	0.559	0.603	0.516
Any pending case	0.620	0.621	0.612	0.620	0.618
Female	0.156	0.156	0.156	0.148	0.168
Black	0.615	0.613	0.608	1.000	0.000
Age	35.2	35.2	35.2	34.6	36.1
Cases	382,635	375,801	245,394	235,465	147,170
Judges	371	362	223	371	371

Notes. Means across cases. Column 1 is the estimation sample; columns 2 and 3 are the cases decided by the judges in the second and third rows of Table D.3; columns 4 and 5 split the estimation sample by race. Misconduct among released is a failure to appear or a rearrest before the case is resolved, among released defendants. A violent top charge is a top arraignment charge in the violent-charge category, felony or misdemeanor. Prior convictions and pending cases are of any class. Section D.5 discusses the comparison.

E Training the Algorithm

This appendix describes the algorithmic risk score M used throughout the paper (Section 3.2). Section E.1 lists the characteristics. Section E.2 describes the estimator, its settings and the calibration step, Section E.3 the cross-fitting procedure, and Section E.4 the performance of the score. Section E.5 describes the models for the alternative outcomes.

E.1 Characteristics

Table E.1 lists the characteristics. All are recorded at arraignment: the defendant's demographics, the category of the top arraignment charge, counts of prior convictions, indicators of pending cases, and the court. The court enters as a categorical variable with 13 levels; the other characteristics enter as indicators, counts or, for age, in years.

E.2 Estimator

The score is a gradient-boosted decision tree classifier (CatBoost) trained on released defendants, for whom the outcome is observed. We minimize the logistic loss with up to 5,000 trees of depth 6, a learning rate of 0.03 and an L_2 penalty of 3 on leaf values, and we weight the two classes equally. Training stops at the number of trees that maximizes the area under the ROC curve on a validation sample held out from training (Section E.3), and the model

Table E.1: Characteristics Used by the Prediction Models

Characteristic	Coding
Demographics	
Age	Years
Female	Indicator
Gender not recorded	Indicator
White	Indicator
Black	Indicator
Hispanic	Indicator
Top arraignment charge	
Felony (class A to E)	Indicator
Violent charge (homicide, rape, robbery, assault, aggravated harassment, strangulation)	Indicator
Robbery	Indicator
Burglary	Indicator
Larceny	Indicator
Other property offense	Indicator
Rape or other sex offense	Indicator
Assault	Indicator
Aggravated harassment or strangulation	Indicator
Criminal possession of a weapon	Indicator
Drug offense	Indicator
Driving while intoxicated	Indicator
Unlicensed operation of a vehicle	Indicator
Criminal contempt	Indicator
Criminal trespass	Indicator
Endangering welfare	Indicator
Hate crime	Indicator
Conspiracy, obstruction or other offense	Indicator
Criminal history	
Prior violent felony convictions	Count, top-coded at 2
Prior other felony convictions	Count, top-coded at 4
Prior misdemeanor convictions	Count, top-coded at 10
Pending violent felony case	Indicator
Pending other felony case	Indicator
Pending misdemeanor case	Indicator
Court	
Court of arraignment	Categorical (13 courts)

Notes. Every prediction model (any pretrial misconduct, failure to appear, failure to appear or violent-felony rearrest) uses the same characteristics. Indicators of the top arraignment charge's category are not exclusive. Conviction counts are top-coded in the source data. No case in the estimation sample has a hate-crime top charge, so that indicator is constant. Section E.1 describes how the characteristics enter.

at that number of trees is kept.

Because the classes are weighted equally, the raw predictions overstate the probability of misconduct. We therefore map them to probabilities by Platt scaling, a logistic regression of the outcome on the log odds of the raw prediction, fit on the same validation sample. The map is increasing, so it leaves the ranking of cases within a fold, and hence the area under the ROC curve, unchanged. The analysis uses the score only through its ranking: the deciles of Sections 4 to 6 and the cutoffs of the algorithm's frontier are formed from the raw predictions, and the scaled predictions are used to assess calibration (Figure E.2, Panel B).

E.3 Cross-Fitting

Every prediction is out of sample. Figure E.1 summarizes the design. The score is constructed in five steps.

1. Split the estimation sample at random by case into two halves of 191,289 and 191,346 cases, subject to every judge and every court appearing in both halves.

2. Within each half, hold out 10 percent of cases as the validation sample, subject to every judge having cases in both the training and the validation parts and every court appearing in the validation part. This leaves 172,155 and 172,211 training cases and 19,134 and 19,135 validation cases.

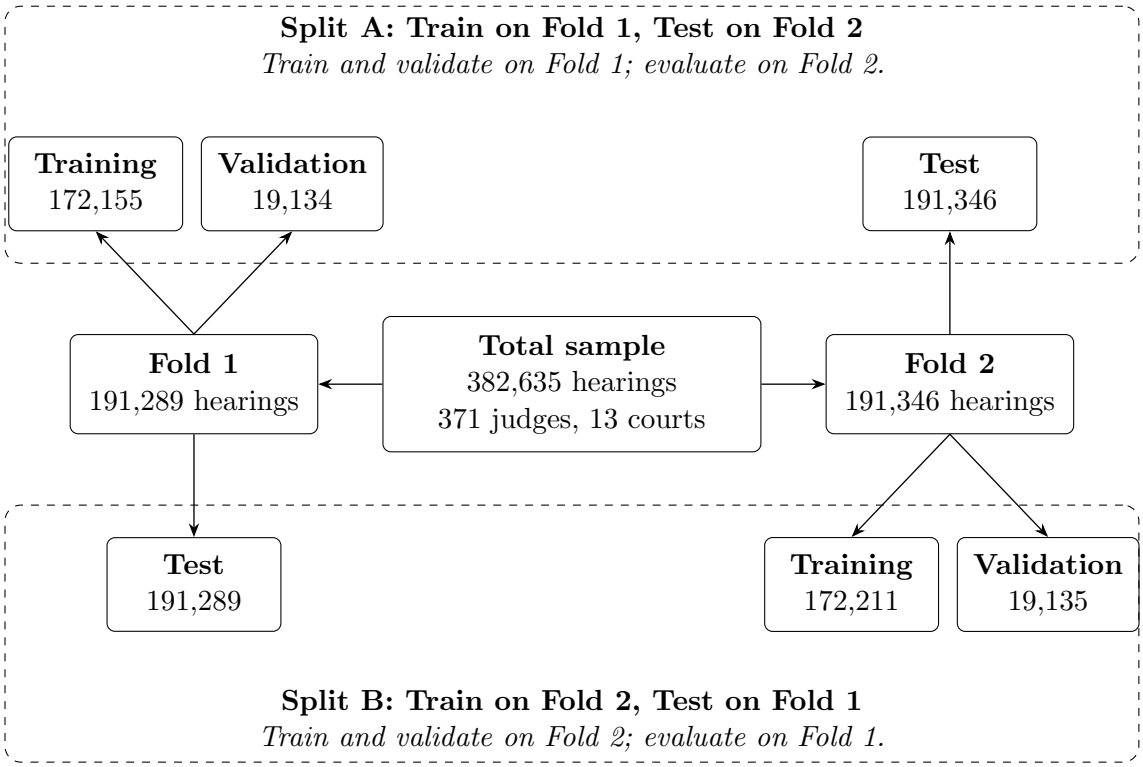
3. On the released defendants in each half's training part, fit the classifier of Section E.2, stopping at the number of trees that maximizes the area under the ROC curve on that half's validation part. The two halves use different random seeds.

4. Fit the Platt scaling map on the released defendants in each half's validation part.

5. Score every case in each half with the model trained on the other half. The score M of a case is this out-of-fold prediction, and the deciles of the score used in Sections 4 to 6 are formed on the full estimation sample.

Coarsening the score into deciles: Sections 4.3.2, 5.3 and 6.2 use the score through its deciles: the prediction M in the framework of Section 2 is the decile of the score, and every bound and test conditions on it. Coarsening keeps the ranking of cases across deciles and discards the variation in the score within a decile. Two consequences follow. First, the algorithm in those sections is the decile rule, whose frontier lies weakly inside the frontier of the continuous score in Figure 3; the comparisons there are therefore with a slightly less informative algorithm than the one in Figure 3. Second, nothing in the bounds or the tests

Figure E.1: Cross-Fitting Design used in Construction of the Algorithmic Risk Score



2-fold cross-split drawn at the case level: in each split, every judge and every court appears in the training, validation, and test sets.

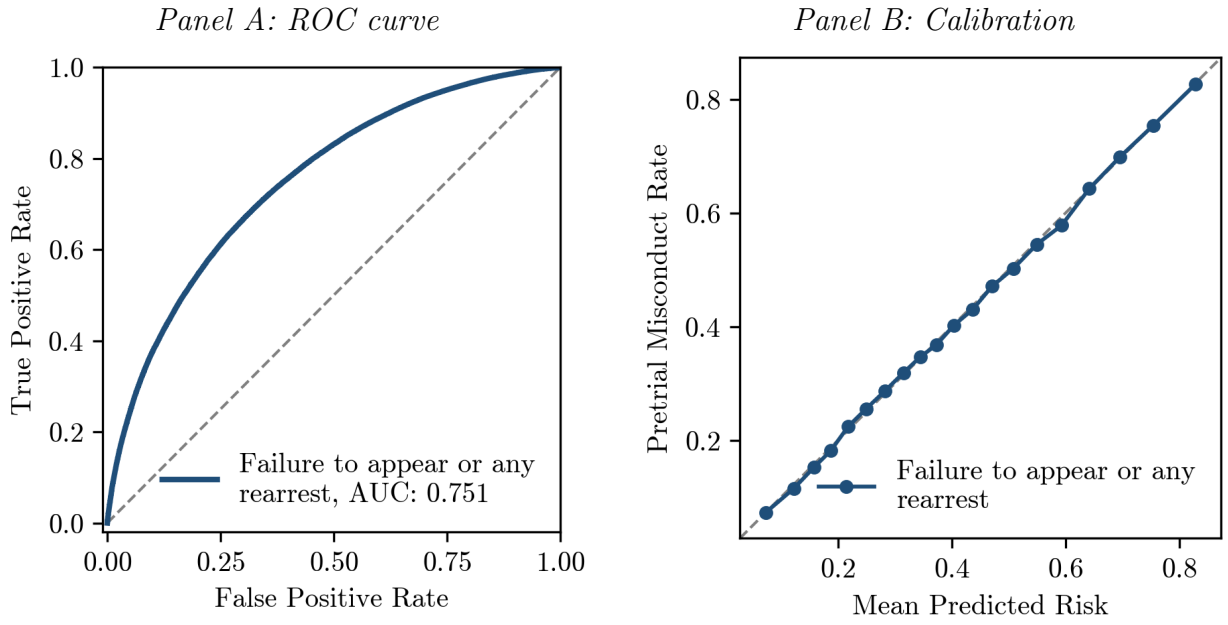
Notes. This figure summarizes the two-fold cross-fitting design. The estimation sample is divided in half at the case level; each half trains and tunes a model that scores the cases of the other half, so that every case's prediction is out of sample. Every judge and every court appears in both halves.

requires the score to be the true probability of misconduct given the characteristics. The test for private information compares the outcome rates of released and detained defendants within each decile, and it remains a valid test of the restrictions it targets for any partition of the cases into deciles.

E.4 Performance

Figure E.2 reports the ROC curve and the calibration of the score among released defendants, pooling the two halves. The area under the ROC curve is 0.751 (0.749 and 0.753 in the two halves). At the cutoff that maximizes the sum of sensitivity and specificity, the score classifies 67.3 percent of released defendants who commit misconduct as high risk and 69.4 percent of those who do not as low risk. After Platt scaling, the misconduct rate in each of 20 equal-size bins of the score is within 1.5 percentage points of the mean prediction; before scaling, the raw predictions exceed the misconduct rate by 7.3 percentage points on average.

Figure E.2: Performance of the Predictive Algorithm on Released Defendants



Notes. Out-of-fold predictions of the misconduct model for released defendants, pooled across the two halves. Panel B plots the misconduct rate against the mean of the scaled prediction within 20 equal-size bins of the score. The solid line in Panel A is the ROC curve and the dashed line random guessing; the points in Panel B are the bins and the dashed line the 45-degree line. Section E.4 discusses the figure.

E.5 Models for the Alternative Outcomes

For each alternative outcome in Appendix I, we train a separate model with the same characteristics, settings, folds and calibration step, on released defendants.

The areas under the ROC curve are 0.764 for failure to appear and 0.743 for failure to appear or violent-felony rearrest (Figure I.1).

F Extrapolation of Misconduct Risk

This appendix gives the details of the extrapolation of misconduct risk across judges described in Section 3.3. Section F.1 restates the construction. Sections F.2 and F.3 describe how it is applied to risk sets of the score and to deciles of the score. Section F.4 describes the bootstrap, Section F.5 reports diagnostics and alternative fits, and Section F.6 reports the evidence on quasi-random assignment. Throughout, a cell is a court and the month of first arraignment (Appendix D.3).

F.1 Construction

Recall regressions (4) and (5). For a population of cases (the full sample, a risk set or a decile), the risk at full release is estimated in four steps.

1. Regress the release indicator C_i and the indicator of release with misconduct C_iY_i on judge indicators and cell indicators, with no intercept and the first cell omitted.

2. Form judge z's strata-adjusted release rate $\hat{S}_{2z} = \hat{\zeta}_z + \sum_s w_s \hat{\gamma}_s^R$ and strata-adjusted rate of release with misconduct $\hat{S}_{1z} = \hat{\alpha}_z + \sum_s w_s \hat{\gamma}_s^A$, where w_s is the share of the population's cases in cell s.

3. Clip $\hat{S}_{2z}$ to $[0, 1]$ and $\hat{S}_{1z}$ to $[0, \hat{S}_{2z}]$. Every judge with cases in the population is kept, including judges whose adjusted release rate is clipped at one.

4. Fit $\hat{S}_{1z}$ on $\hat{S}_{2z}$ across judges by least squares weighted by each judge's number of cases in the population, and read off the fitted value at $\hat{S}_{2z} = 1$.

Because a judge who releases every case observes misconduct for all of them, the fitted value at full release estimates the population's risk $P(Y^* = 1)$ under the assumption stated in Section F.6. For the full sample, the fitted line has an intercept and its value at one is 42.6 percent. As a check, we also report a local linear fit at one with a Gaussian kernel in the release rate, bandwidth 0.1, and the same case weights; it gives 43.7 percent.

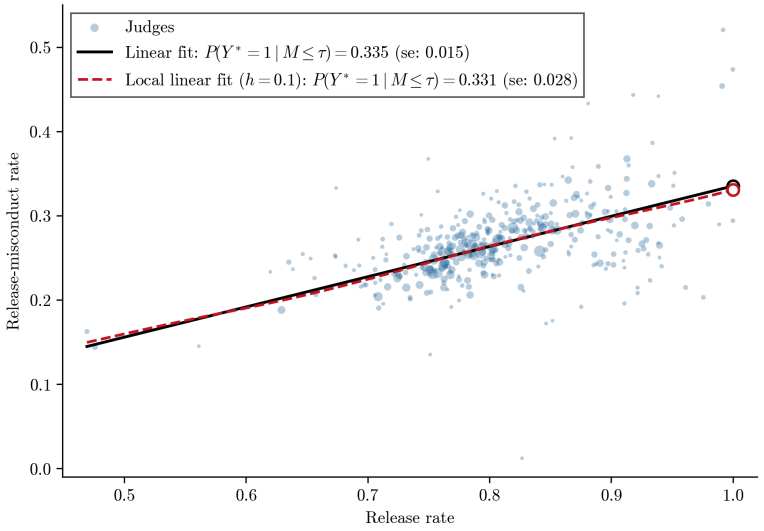
F.2 Risk Sets of the Score

The algorithm's frontier in Figure 3 requires $P(Y^* = 1, M \leq t)$ for cutoffs t of the score. For each cutoff we apply the construction of Section F.1 to the cases with $M \leq t$, using that population's own mix of cells, and multiply the fitted value at full release by the share of cases with $M \leq t$. The cutoffs are the quantiles of the score at 40 release rates evenly spaced

from just below the release rate of the least lenient judge (0.47) to 0.995, together with release rates of 0.05, 0.10, and so on below that range. The frontier joins the resulting points with the origin. Below the least lenient judge no judge's release rate is near the evaluation point, so that part of the frontier rests on extrapolation alone; none of our comparisons uses it.

Figure F.1 shows the construction at the cutoff τ at which the algorithm releases as many defendants as the judges, $P(M \leq \tau) = 0.796$. The fitted value at full release is 33.5 percent (local linear, 33.1 percent).

Figure F.1: Extrapolating Misconduct Risk Across Judges: Matched Decision Rule



Notes. As Figure 1, for the risk set $\{M \leq \tau\}$ at the cutoff τ at which the algorithm releases as many defendants as the judges. Each point is one judge's strata-adjusted release rate and rate of release with misconduct, sized by caseload. The solid line is the caseload-weighted linear fit and the dashed line the local linear fit (Gaussian kernel, bandwidth 0.1); the open circles mark their values at full release. Section F.2 discusses the figure.

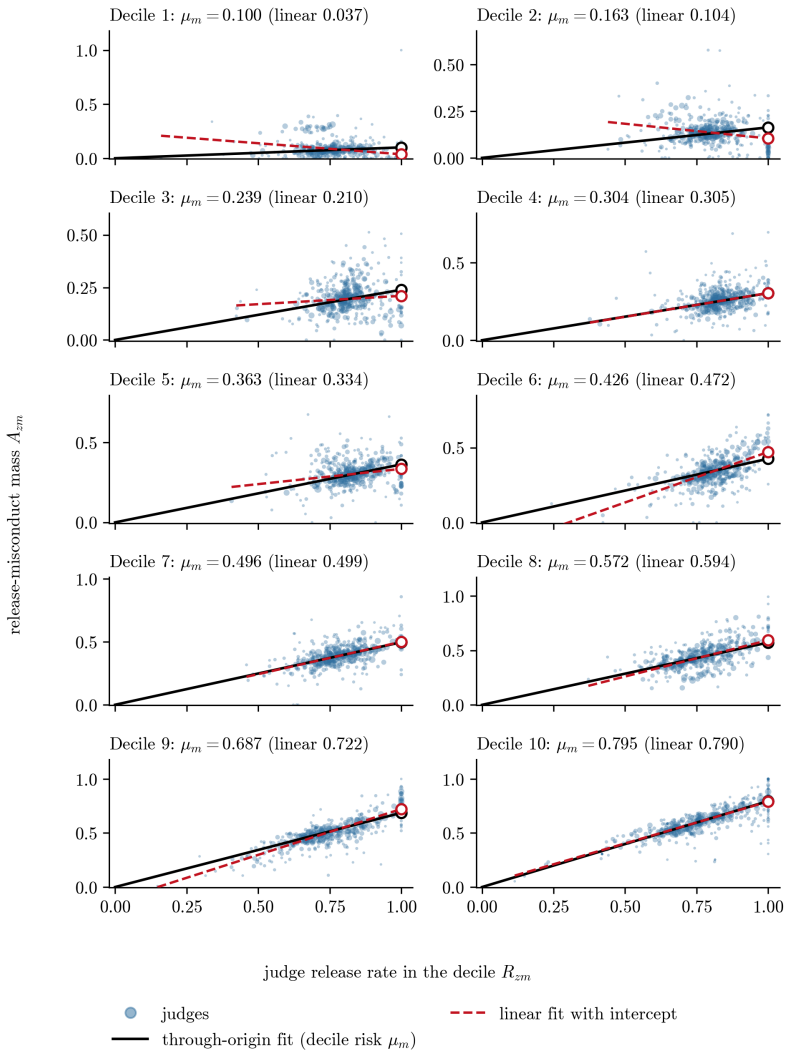
F.3 Deciles of the Score

The bounds in Sections 4.3.2, 5.3.2 and 6.2 require the risk within each decile of the score, $\mu(m) = P(Y^* = 1 \mid M = m)$. Within each decile, we estimate regressions (4) and (5) on the decile's cases, among judges with at least 50 cases in the estimation sample, and form each judge's strata-adjusted rates at the decile's own mix of cells. We then fit the rate of release with misconduct on the release rate across judges through the origin, weighting by each judge's cases in the decile:

$$\hat{\mu}(m) = \min\left\{1, \frac{\sum_z n_{zm} \hat{S}_{2zm} \hat{S}_{1zm}}{\sum_z n_{zm} \hat{S}_{2zm}^2}\right\}. \tag{A19}$$

The fit through the origin imposes that a judge who releases no one has no released misconduct, which holds by construction. Figure F.2 shows the ten fits. The estimated risks are 10.0 percent in the first decile (standard error 1.6) and 79.5 percent in the tenth (standard error 0.4).

Figure F.2: Extrapolating Misconduct Risk Within Deciles of the Algorithm's Score



Notes. One panel per decile of the score. Each point is one judge's strata-adjusted release rate and rate of release with misconduct on the decile's cases, sized by the judge's cases in the decile. The solid line is the caseload-weighted fit through the origin, equation (A19), and the dashed line the caseload-weighted linear fit with an intercept. The open circles mark their values at full release. Section F.3 discusses the figure.

F.4 Inference

We measure sampling uncertainty with the Bayesian bootstrap (Rubin, 1981; Chamberlain and Imbens, 2003). For each replication:

1. Draw a weight for every case from the standard exponential distribution.

2. Repeat steps 1 to 4 of Section F.1 with every regression and every fit across judges weighted by these case weights.

3. Record the values at full release and every quantity built from them.

Standard errors are standard deviations across replications, and intervals are their 5th and 95th percentiles. The full sample and the matched decision rule use 500 replications. The deciles use the 500 replications of the private information tests (Section 5.3.1), which give the intervals in Figure F.3. The band in Figure F.4 comes from 500 replications of the frontier, each repeating the construction of Section F.2 at every cutoff.

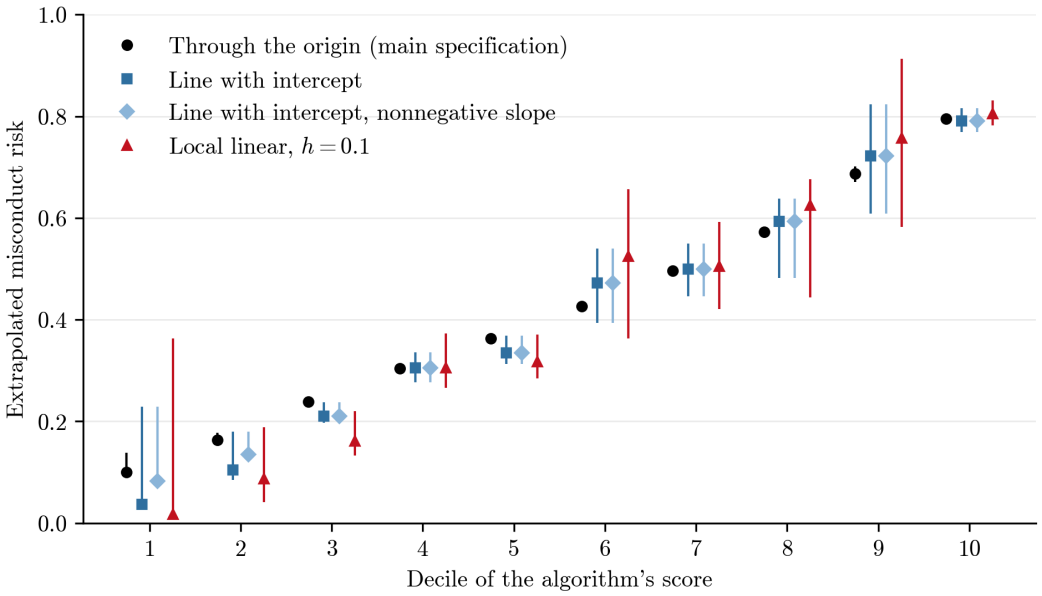
F.5 Diagnostics and Alternative Fits

Alternative fits within deciles: Figure F.3 reports each decile's risk under four fits: through the origin (equation (A19)), a line with an intercept read at full release, the same line constrained to a nonnegative slope, and the local linear fit. The fitted slope of the line with an intercept is negative in deciles 1 and 2, which the model rules out; constraining the slope to be nonnegative sets it to zero there. In the first decile, the line with an intercept gives 3.7 percent at full release, the constrained line 8.2 percent and the fit through the origin 10.0 percent. The fits with an intercept are imprecise: across the ten deciles, the 90 percent intervals of the line with an intercept and of the local linear fit are 4 to 35 percentage points wide, and these fits differ from the fit through the origin by up to 9.8 percentage points (decile 6). Their intervals contain the fit through the origin in 18 of the 20 cases.

Uncertainty in the frontier: Figure F.4 adds to the frontier of Figure 3 a pointwise 90 percent band over the judges' release rates, from 0.47 to 0.99, and the frontier obtained with the local linear fit at every cutoff. The band is 2.8 to 4.9 percentage points of release-misconduct mass wide, and the local linear frontier stays within 1.5 percentage points of the linear one over the judges' release rates.

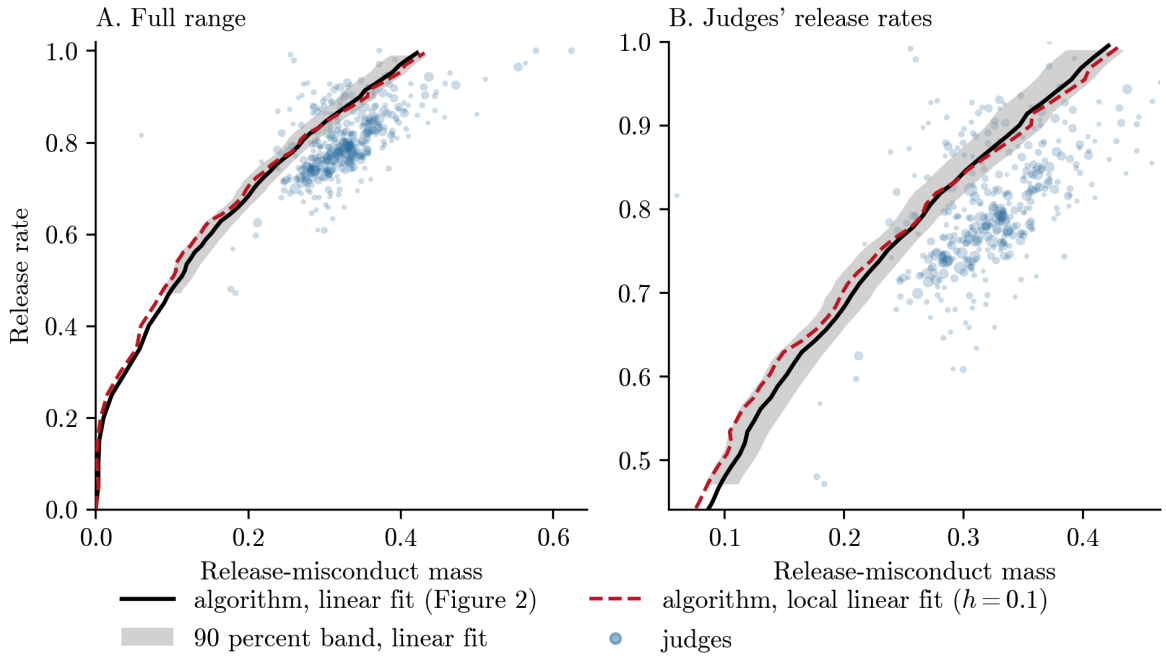
Bandwidth: For the full sample, the local linear value at full release is 49.2 percent with a bandwidth of 0.03, 43.7 percent with 0.1 and 42.8 percent with 0.3, against 42.6 percent for the linear fit. At small bandwidths the fit rests on the few judges who release nearly every case. At the matched cutoff the corresponding values are 39.9, 33.1 and 33.5 percent, against 33.5 percent.

Figure F.3: Extrapolated Risk by Decile Under Alternative Fits



Notes. Each decile's risk from four caseload-weighted fits of the rate of release with misconduct on the release rate across judges: through the origin (equation (A19)), a line with an intercept, the same line with its slope constrained to be nonnegative (when the fitted slope is negative it is set to zero and the value is the caseload-weighted mean rate), and a local linear fit with bandwidth 0.1. Markers are point estimates and whiskers the 5th to 95th percentiles over 500 Bayesian bootstrap replications. Section F.5 discusses the figure.

Figure F.4: The Algorithm's Frontier and Its Uncertainty



Notes. The solid line is the algorithm's frontier in Figure 3. The shaded band is a pointwise 90 percent interval over 500 Bayesian bootstrap replications of the frontier, drawn over the judges' release rates. The dashed line is the frontier with a local linear fit (bandwidth 0.1) in place of the linear fit at each cutoff. Points are judges, sized by caseload. Panel B enlarges the judges' release rates. Section F.5 discusses the figure.

Leaving out one court at a time: Release rates differ across courts (Table 1), so we re-estimate the full-sample risk and the ten decile risks dropping each of the 13 courts in turn. The full-sample risk moves by at most 2.5 percentage points (dropping Nassau District Court) and no decile risk moves by more than 1.4 percentage points.

F.6 Quasi-Random Assignment

The extrapolation rests on the following assumption, stated in Section 3.3.

Assumption A1 (Quasi-random assignment within cells)**.** Within each cell S, the assigned judge Z is independent of the algorithm's score and of the defendant's potential misconduct, $Z \perp\!\!\!\perp (M, Y^*) \mid S$.

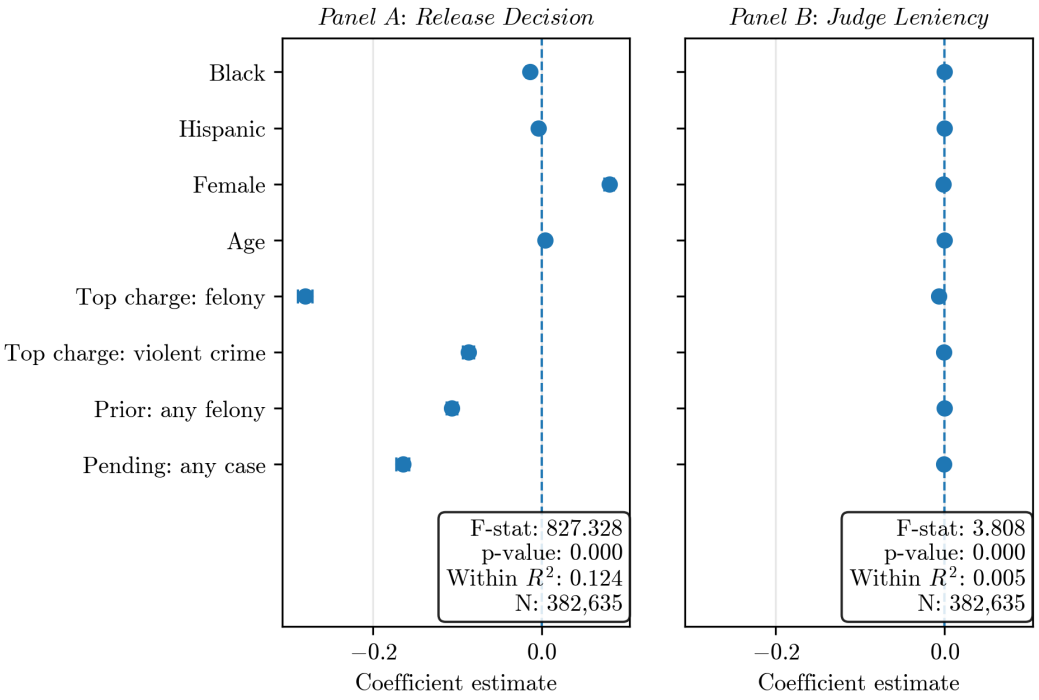
Each cell contained at least 100 cases and at least 2 judges when it was selected (Appendix D.2), and every cell still contains at least two judges after the requirement on judges. Figure F.5 reports the balance test of Section 3.3. Panel A regresses the release decision on the case characteristics with cell fixed effects; Panel B regresses the assigned judge's leave-out release rate on the same characteristics and fixed effects. Standard errors are clustered by judge. The characteristics predict the release decision strongly (joint F-statistic 827.3, within R^2 0.124). They predict the judge's leave-out release rate much less: the joint F-statistic is 3.81 with a p-value of 0.000, so the test rejects at the 5 percent level; the within R^2 is 0.005, and the largest coefficient in absolute value (Top charge: felony) is -0.0071, or 0.71 percentage points of the leave-out release rate. Table F.1 repeats both regressions within each group of defendants used in Appendix J. The joint test on the leave-out release rate rejects at the 5 percent level in all four groups; in every group the release decision is predicted far more strongly. The extrapolation's identifying variation is the judge's release rate within cells, of which the characteristics explain a within R^2 of 0.005 in the full sample.

Table F.1: Balance of Case Characteristics Across Judges, by Group

			Release decision		Judge's leave-out release rate	
	Cases	Judges	F	p-value	F	p-value
All defendants	382,635	371	827.3	<0.001	3.81	<0.001
Black defendants	235,465	371	875.7	<0.001	4.31	<0.001
White defendants	147,170	371	545.9	<0.001	3.48	0.001
Male defendants	322,290	371	935.2	<0.001	4.34	<0.001
Female defendants	59,658	371	182.2	<0.001	2.48	0.017

Notes. Joint F-tests that the coefficients on the case characteristics of Figure F.5 are zero, in regressions of the release decision and of the assigned judge's leave-out release rate on the characteristics with cell fixed effects, estimated within each group of defendants. Standard errors are clustered by judge. Characteristics that are constant within a group are dropped. Section F.6 discusses the table.

Figure F.5: Balance of Case Characteristics Across Judges



Notes. Coefficients from regressions of the release decision (Panel A) and of the assigned judge's leave-out release rate (Panel B) on case characteristics, with cell fixed effects. Points are coefficients and bars 95 percent confidence intervals, with standard errors clustered by judge; the boxes report the joint F-test, its p-value, the within R^2 and the number of cases. Section F.6 discusses the figure.

G Empirical Bayes

This appendix describes the empirical Bayes procedure behind the shares of judges reported in Section 4.3.1: the share whose realized choices lie inside the algorithm's frontier at their own release rates, and the share below the random-release line. Section G.1 defines the judge-level estimates and their sampling variances. Section G.2 describes the prior and the posterior shares, and Section G.3 compares the empirical Bayes shares with the raw shares.

G.1 Judge-Level Estimates

Recall from Appendix F.1 that judge z's strata-adjusted release rate is $\hat{S}_{2z}$ and the judge's strata-adjusted rate of release with misconduct is $\hat{S}_{1z}$. The judge's misconduct rate among released defendants is $\hat{\rho}_z = \hat{S}_{1z}/\hat{S}_{2z}$. We classify each judge by two gaps, each positive when the judge's released defendants are riskier than the benchmark:

$$\hat{d}_z^{\text{alg}} = \hat{\rho}_z - \hat{a}(\hat{S}_{2z}), \qquad \hat{d}_z^{\text{rand}} = \hat{\rho}_z - \hat{\mu}, \tag{A20}$$

where $\hat{a}(r)$ is the misconduct rate among the defendants the algorithm releases at release rate r, read off the frontier of Figure 3, and $\hat{\mu}$ is the frontier's value at full release, the misconduct rate of a rule that releases at random. A judge is inside the algorithm's frontier if $d_z^{\text{alg}} > 0$ and below the random-release line if $d_z^{\text{rand}} > 0$.

The sampling variance of each gap comes from 500 replications of the Bayesian bootstrap (Appendix F.4). In each replication, we redraw the case weights and recompute the judge's strata-adjusted rates, the frontier at every cutoff and its value at full release, and then both gaps. The standard error s_z of a gap is its standard deviation across replications. It therefore reflects the sampling error in the judge's own rates, in the frontier and in $\hat{\mu}$, and their covariance.

G.2 Prior and Posterior Shares

Following Angelova, Dobbie and Yang (2023), we treat each true gap d_z as a draw from a normal distribution $N(m, \tau^2)$ and each estimate as $\hat{d}_z \mid d_z \sim N(d_z, s_z^2)$. For each of the two gaps, the share is computed in three steps.

1. Estimate m and τ by maximum likelihood from the marginal distribution $\hat{d}_z \sim N(m, \tau^2 + s_z^2)$ across the 371 judges.

2. For each judge, form the posterior distribution of d_z: normal with mean $m + \lambda_z(\hat{d}_z - m)$ and variance $\lambda_z s_z^2$, where $\lambda_z = \tau^2/(\tau^2 + s_z^2)$.

3. Average across judges the posterior probability that $d_z > 0$.

For the gap with the algorithm, $\hat{m} = 7.4$ percentage points and $\hat{\tau} = 4.2$ percentage points; for the gap with random release, $\hat{m} = -0.7$ and $\hat{\tau} = 4.1$. Figure G.1 shows the estimates, the fitted distributions and the posterior means. The average shrinkage factor λ_z is 0.66 for the first gap and 0.66 for the second. Across judges, the median bootstrap standard error is 2.9 percentage points for the gap with the algorithm and 2.8 for the gap with random release, and the standard errors range from 0.9 to 9.6. The posterior probability that a judge is inside the algorithm's frontier has a median of 0.999; it exceeds 0.95 for 83.3 percent of judges and is below 0.05 for 0.8 percent. The posterior probability that a judge is below the random-release line has a median of 0.407; it exceeds 0.95 for 7.0 percent of judges and is below 0.05 for 15.9 percent. The shares are averages of posterior probabilities, not counts of judges classified at a stated error rate, and we do not report uncertainty in the shares themselves.

G.3 Comparison with Raw Shares

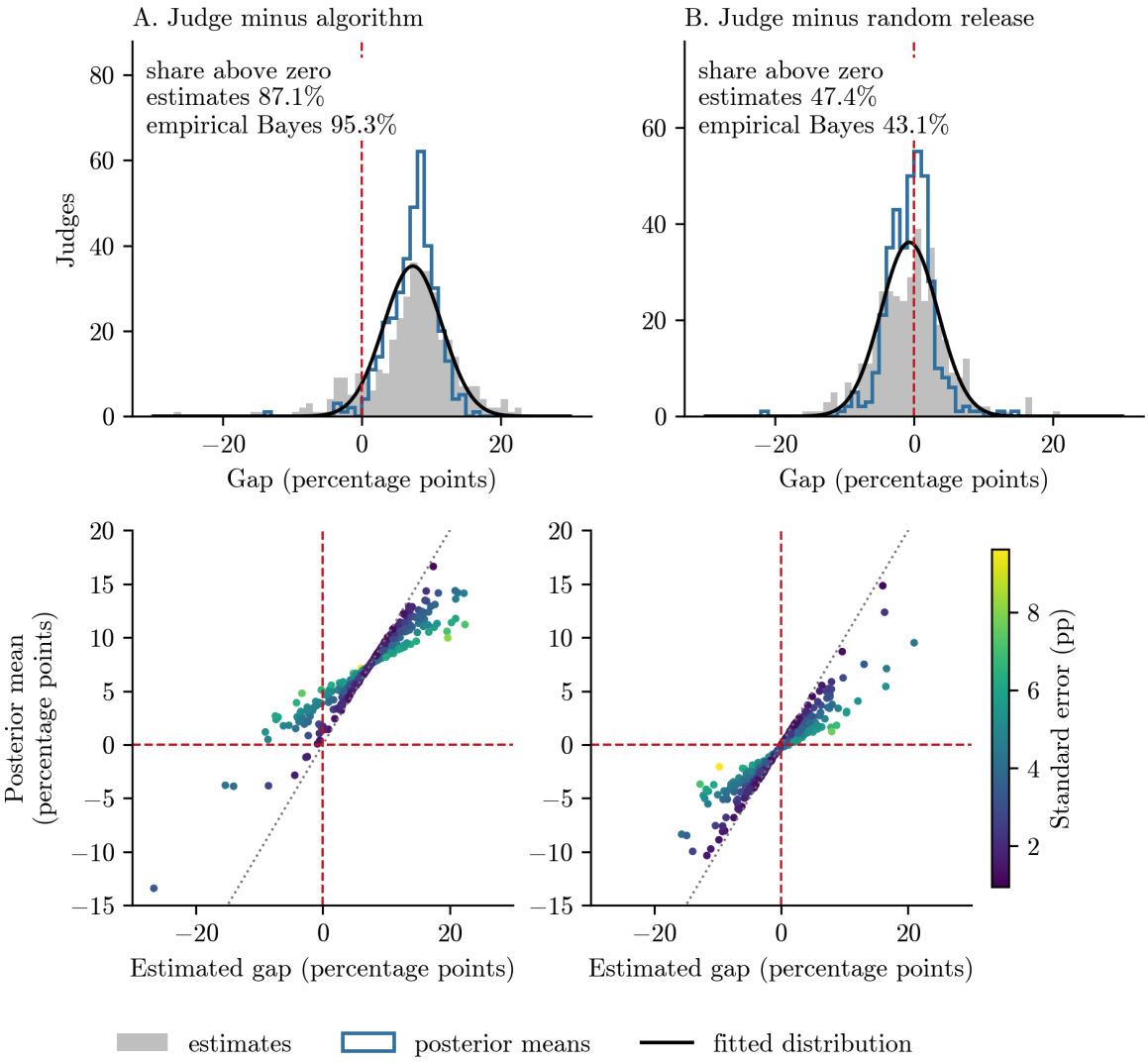
Table G.1 compares the shares computed from the raw estimates with the empirical Bayes shares under two sets of standard errors: binomial standard errors on each judge's released caseload, which treat the frontier and $\hat{\mu}$ as known, and the bootstrap standard errors described above, which we use in the main text. The raw estimates place 87.1 percent of judges inside the algorithm's frontier and 47.4 percent below the random-release line. Shrinkage moves both shares away from one half, to 95.3 and 43.1 percent, because the fitted distribution of the gap with the algorithm is centered well above zero and that of the gap with random release below it. The binomial standard errors understate the sampling variance, shrink less, and give 89.9 and 44.6 percent.

Table G.1: Share of Judges Above Each Benchmark: Raw and Empirical Bayes Estimates

		Empirical Bayes	
Share of judges	Raw estimates	Binomial standard errors	Bootstrap standard errors
Inside the algorithm's frontier	87.1	89.9	95.3
Below random release	47.4	44.6	43.1
Fitted distribution of the gap			
Standard deviation, inside the algorithm's frontier		5.3	4.2
Standard deviation, below random release		4.6	4.1

Notes. Percent of the 371 judges whose misconduct rate among released defendants exceeds that of the algorithm at the judge's release rate (first row) and that of random release (second row). Raw estimates count point estimates above zero. Empirical Bayes shares average the posterior probability of a positive gap under a normal prior fitted by maximum likelihood, with binomial standard errors on the judge's released caseload or with Bayesian bootstrap standard errors (500 replications). The last two rows report the fitted standard deviation $\hat{\tau}$ of the true gap, in percentage points. Section G.3 discusses the table.

Figure G.1: Estimated and Shrunk Gaps



Notes. Panel A: each judge's misconduct rate among released defendants minus the algorithm's at the judge's release rate. Panel B: minus the misconduct rate under random release. Top row: histogram of the estimated gaps, the normal distribution fitted by maximum likelihood, and the histogram of posterior means. Bottom row: posterior mean against the estimate, colored by the bootstrap standard error; the dotted line is the 45-degree line and the dashed red lines mark a gap of zero. 371 judges. Section G.2 discusses the figure.

H Testing for Private Information

This appendix describes the test of Section 5.3.1. By Lemma 2(i), judge z has private information at decile m of predicted risk if and only if the judge's choices and misconduct are dependent within that decile, so we test the null

$$C \perp\!\!\!\perp Y^* \mid M = m,\ Z = z$$

for each judge and decile.

Moments: For judge z and decile m, let $S_{2z}(m)$ and $S_{1z}(m)$ denote the judge's strata-adjusted release rate and rate of release with misconduct among the decile's cases, constructed as in Section 3.3. Within each decile, we regress C_i, and separately $C_i Y_i$, on a full set of judge dummies with no intercept and on stratum dummies with the first stratum omitted, and we evaluate each judge's fitted value at the decile's mix of strata, where a stratum is a cell of Appendix D.3. We clip $S_{2z}(m)$ to $[0, 1]$ and $S_{1z}(m)$ to $[0, S_{2z}(m)]$. We call a pair of a judge and a decile a cell. There are 3,701 cells rather than 3,710, because fewer than 11 judges have no cases in the top decile.

Benchmark and residual: Under the null, the two moments factor as $S_{1z}(m) = S_{2z}(m)\,\mu(m)$, where $\mu(m)$ is the decile's risk. We estimate the benchmark for judge z from the decile's other judges, by a regression through the origin of their rates of release with misconduct on their release rates, with equal weight on each judge,

$$\widehat{\mu}_{-z}(m) := \frac{\sum_{z' \neq z} S_{2z'}(m)\, S_{1z'}(m)}{\sum_{z' \neq z} S_{2z'}(m)^2},$$

clipped to $[0, 1]$. Leaving out the tested judge keeps the benchmark independent of that judge's moments. The regression goes through the origin because the null implies no intercept, so the slope is also the fitted value at a release rate of one, the risk faced by a judge who releases every defendant. The residual for each cell is

$$D_z(m) := S_{1z}(m) - S_{2z}(m)\,\widehat{\mu}_{-z}(m),$$

which is zero under the null. A positive residual means that the defendants the judge releases in the decile are riskier than the decile as a whole, and a negative residual means that they are less risky.

Bootstrap: We use a Bayesian bootstrap with 500 draws. Each draw assigns independent exponential weights to the cases and recomputes every step, including the regressions in each decile, the clipping, and the benchmarks. For each cell, $\sigma_z(m)$ is the standard deviation of $D_z(m)$ across draws, and $t_z(m) := D_z(m)/\sigma_z(m)$ is the studentized residual. For each judge, $\widehat{\Omega}_z$ is the correlation matrix of the judge's residuals across deciles, computed across draws. Because one set of weights generates all of a judge's cells in a draw, $\widehat{\Omega}_z$ captures the dependence across deciles that comes from shared strata and shared benchmarks. Every decile in which the judge has cases is included.

Tests: The test within a cell is two-sided, with p-value $2\,\Phi(-|t_z(m)|)$. The test for each judge pools the judge's deciles. Its statistic is $\max_m|t_z(m)|$, and its p-value is the share of 100,000 draws of $\xi \sim N(0, \widehat{\Omega}_z)$ for which $\max_m|\xi_m|$ exceeds the statistic.

Multiple testing: The tests for judges and the tests for cells form two separate families, with 371 and 3,701 hypotheses. Within each family, our headline adjustment is the Benjamini-Hochberg procedure at 5 percent (Benjamini and Hochberg, 1995), which controls the false discovery rate. We also report Holm's step-down procedure at 5 percent (Holm, 1979), which controls the probability of any false rejection. It compares the i-th smallest of the family's n p-values to $0.05/(n-i+1)$ and stops at the first that fails. The main text reports the shares of judges whose test rejects, and the tests for cells describe where in the distribution of predicted risk each judge's private information lies.

I Results Under Alternative Definitions of Pretrial Misconduct

This appendix repeats the analysis of Sections 3.3, 4.3.1 and 5.3.1 for two alternative definitions of pretrial misconduct. Section I.1 describes the outcomes and the corresponding algorithms. Section I.2 asks whether the algorithm outperforms the judges under each outcome, and Section I.3 whether the judges have private information about it. Table I.1 collects the headline numbers for the three outcomes.

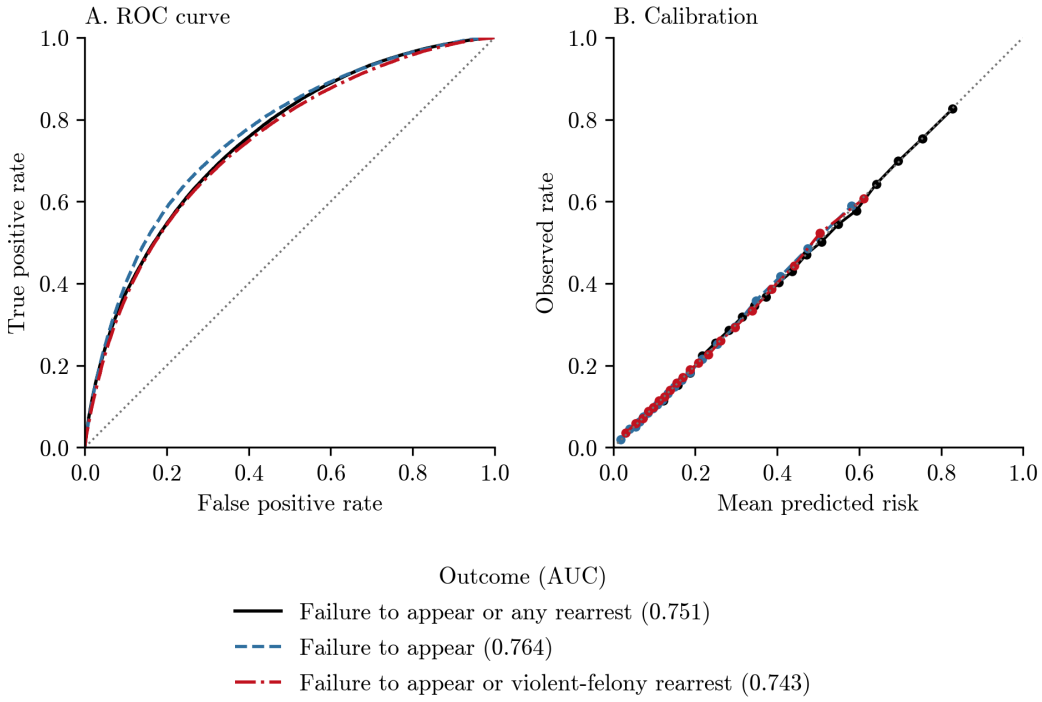
I.1 Outcomes and Algorithms

The main text measures pretrial misconduct as a failure to appear or any rearrest before the case is resolved, the outcome studied in the literature on pretrial release (Kleinberg et al., 2017; Arnold, Dobbie and Hull, 2022; Rambachan, 2024); 40.8 percent of released defendants commit it. We consider two narrower definitions. The first counts only failures to appear, the outcome the statute directs judges to weigh (19.5 percent of released defendants). The

second counts failures to appear and rearrests for violent felony offenses (22.6 percent), and so drops released defendants whose only misconduct is a rearrest for another offense, 18.2 of the 40.8 percent (Table 1).

For each outcome, we use the algorithm trained to predict that outcome (Appendix E.5), so the score and its deciles change with the outcome; every other step of the analysis is unchanged. Figure I.1 reports the performance of the three algorithms. The algorithm for failure to appear is slightly more accurate than the main one (AUC 0.764 against 0.751) and the algorithm for failure to appear or violent-felony rearrest slightly less (0.743).

Figure I.1: Performance of the Algorithm, by Outcome



Notes. Out-of-fold predictions for released defendants, pooled across the two folds, for the algorithm trained on each outcome: failure to appear or any rearrest (solid black), failure to appear (dashed blue), and failure to appear or violent-felony rearrest (dash-dotted red). Panel A: ROC curves, with the area under each curve in the legend. Panel B: outcome rate against mean prediction after Platt scaling, within 20 equal-size bins of each score. The dotted lines are random guessing (Panel A) and the 45-degree line (Panel B). Section I.1 discusses the figure.

I.2 Comparing Judges and the Algorithm

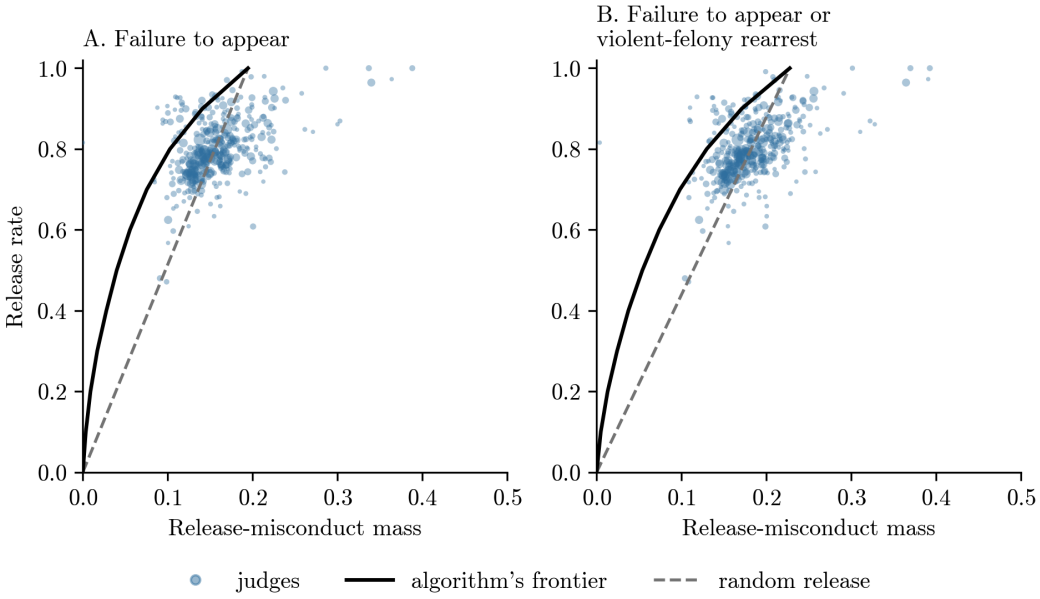
How risky are defendants under each outcome? The extrapolation of Appendix F gives a risk at full release of 21.1 percent for failure to appear and 24.3 percent for failure to appear or violent-felony rearrest, against 42.6 percent for the main outcome (Table I.1). The local linear fit gives 23.2 and 26.2 percent. Among the defendants the algorithm would

release at the judges' release rate, the risks are 13.0 and 16.5 percent, against 33.5 percent.

Does the algorithm outperform the judges? Figure I.2 repeats Figure 3 for the two alternative outcomes. Under the two alternatives, 91.6 and 92.2 percent of judges lie inside the algorithm's frontier at their release rates, against 87.1 percent for the main outcome. These are shares of point estimates; Appendix G shows that for the main outcome empirical Bayes shrinkage raises the share inside the frontier from 87.1 to 95.3 percent.

Do judges do worse than random release? Under the two alternatives, 46.9 and 47.2 percent of judges lie below the random-release line (point estimates), against 47.4 percent for the main outcome.

Figure I.2: The Common Comparison, by Outcome



Notes. As Figure 3, for failure to appear (Panel A) and for failure to appear or violent-felony rearrest (Panel B), with the algorithm trained on each outcome. Each point is one judge's strata-adjusted rate of release with the outcome and release rate, sized by caseload. The solid line is the algorithm's frontier, extrapolated as in Appendix F.2, and the dashed line is random release. Below the least lenient judge's release rate the frontier rests on extrapolation alone. Section I.2 discusses the figure.

I.3 Private Information

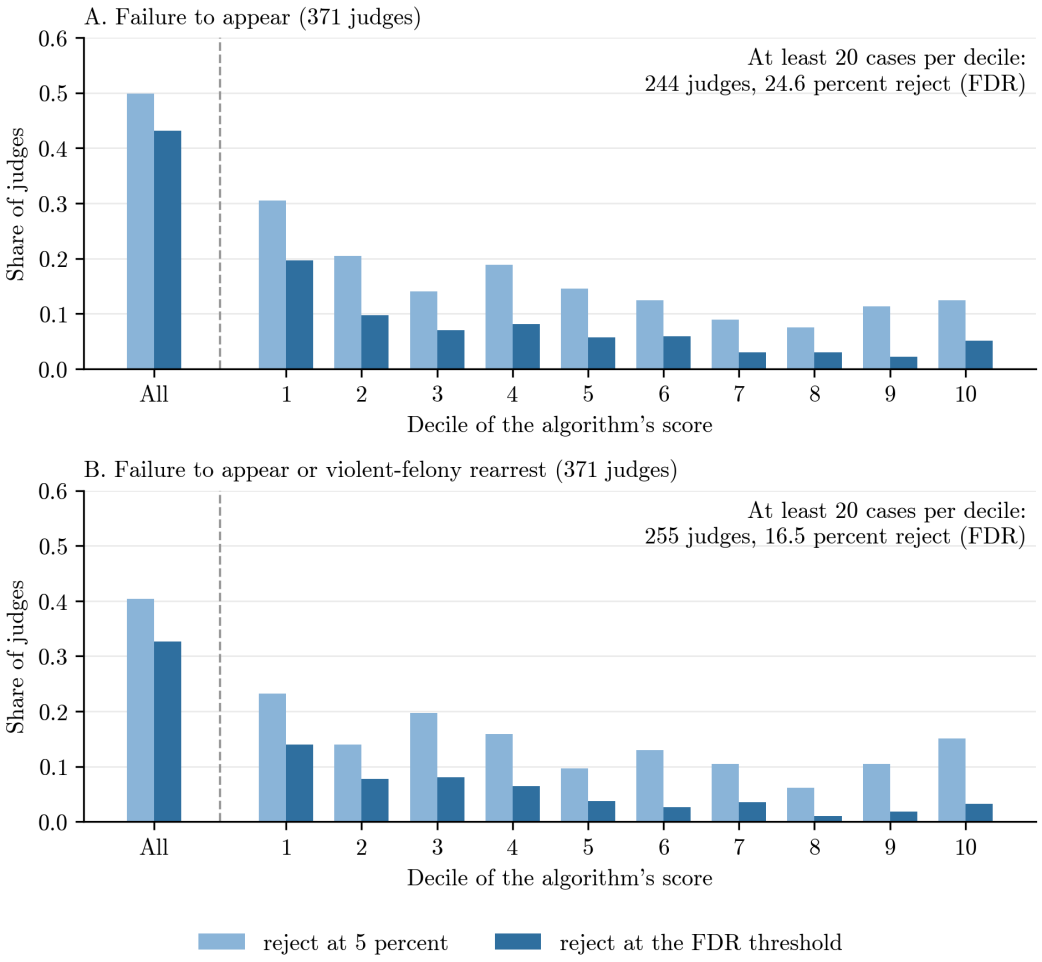
Do judges have private information? Figure I.3 repeats Figure 6. We apply the test of Section 5.3.1 without change: the Bayesian bootstrap with 500 replications per outcome and the same random seeds, a two-sided test within each decile, a test across deciles for each judge, and the same adjustments across judges. Both alternatives show more evidence of

private information than the main outcome. Controlling the false discovery rate at 5 percent, 43.1 percent of judges reject the null of no private information about failure to appear and 32.6 percent about failure to appear or violent-felony rearrest, against 25.1 percent for the main outcome. Controlling the family-wise error rate at 5 percent with Holm's procedure, the shares are 19.4 and 19.7 percent, against 17.0 percent; with probability at least 0.95, every judge in these counts has private information, so at least 19.4 and 19.7 percent of judges do.

Is the evidence driven by small cells? As for the main outcome, many rejections come from small judge-decile cells, often cells with no misconduct among released defendants. Among judges with at least 20 cases in every decile, the false discovery rate shares are 24.6 percent of 244 judges and 16.5 percent of 255 judges, and the Holm shares 9.8 and 9.4 percent, against 5.4 and 2.7 percent of 223 judges for the main outcome.

Are the judges calibrated? The test of conditional calibration rejects for fewer judges under the alternatives: 1.3 and 1.9 percent at the false discovery rate threshold, against 7.5 percent for the main outcome.

Figure I.3: Share of Judges With Evidence of Private Information, by Outcome



Notes. As Figure 6, for failure to appear (Panel A) and for failure to appear or violent-felony rearrest (Panel B), with the algorithm and score deciles trained on each outcome and the outcome's risks extrapolated as in Appendix F.3. Light bars: share of judges rejecting independence between their choices and the outcome at 5 percent; dark bars: after the false discovery rate adjustment. The left pair pools across deciles; the others are within each decile of the outcome's score. The text in each panel reports the share rejecting at the false discovery rate threshold among judges with at least 20 cases in every decile, with the adjustment recomputed among them. Bayesian bootstrap with 500 replications. Section I.3 discusses the figure.

Table I.1: Main Results, by Outcome

	Any misconduct	Failure to appear	Failure to appear or violent-felony rearrest
Cases	382,635	382,635	382,635
Release rate	79.6	79.6	79.6
Outcome rate among released defendants	40.8	19.5	22.6
Extrapolated risk, linear fit	42.6	21.1	24.3
Extrapolated risk, local linear fit	43.7	23.2	26.2
Extrapolated risk at the matched cutoff	33.5	13.0	16.5
Judges inside the algorithm's frontier (raw)	87.1	91.6	92.2
Judges below random release (raw)	47.4	46.9	47.2
Judges tested for private information	371	371	371
Reject no private information (FDR)	25.1	43.1	32.6
Reject no private information (Holm)	17.0	19.4	19.7
Judges with at least 20 cases in every decile	223	244	255
Of these, reject no private information (FDR)	5.4	24.6	16.5
Of these, reject no private information (Holm)	2.7	9.8	9.4
Reject calibration (FDR)	7.5	1.3	1.9

Notes. Percent, except counts of cases and judges. Each column uses the algorithm trained on that outcome. Extrapolated risks are the fitted values at full release of Appendix F.1; the matched cutoff is the score at which the algorithm releases as many defendants as the judges. Shares of judges inside the frontier and below random release are shares of point estimates. The private information and calibration tests are those of Section 5.3.1. FDR rows control the false discovery rate at 5 percent across judges; Holm rows control the family-wise error rate at 5 percent. Section I discusses the table.

J Results by Race and Gender

This appendix reports the analysis of Sections 3.3, 4.3.1 and 5.3.1 separately for Black and white defendants and for male and female defendants. As noted in Section 2.1, our framework can be applied conditional on coarse groups of individuals, in which case it concerns the decision maker's choices across cases within a group. Section J.1 describes how the analysis is conducted within each group. Section J.2 asks whether the algorithm outperforms the judges within each group, and Section J.3 whether the judges have private information. Table J.1 collects the headline numbers.

J.1 Analysis Within Groups

We keep the algorithm of Appendix E, trained on all released defendants, and the deciles of its score in the full estimation sample, so that a decile refers to the same range of predicted risk in every group. We then restrict the cases to one group at a time and repeat every subsequent step on that group's cases alone: the strata-adjusted judge rates, the extrapolation of the risk at full release, the algorithm's frontier and the private information test. Male defendants are those whose recorded gender is male; the 687 cases (0.2 percent) with no recorded gender are excluded from both gender groups. The algorithm uses race and gender as inputs (Table E.1).

The groups differ in size and composition. Black defendants account for 235,465 cases and white defendants for 147,170; male defendants for 322,290 and female defendants for 59,658 (Table J.1). Table D.4 compares the characteristics of Black and white defendants, and Table F.1 reports the balance test within each group. All 371 judges decide cases in every group, but female defendants average 161 cases per judge, so many judge-decile cells are small. The private information test uses judges with at least 50 cases in the group: 367 for Black defendants, 356 for white, 371 for male and 274 for female defendants.

J.2 Comparing Judges and the Algorithm

How risky are defendants in each group? The extrapolated risk at full release is 40.3 percent for Black defendants and 44.2 percent for white defendants, and 42.8 percent for male and 38.8 percent for female defendants, against 42.6 percent for all defendants. The local linear fits give 39.1, 46.8, 43.1 and 39.7 percent. Judges release 77.8 percent of Black defendants and 82.4 percent of white defendants, and 77.8 percent of male and 88.9 percent of female defendants.

Does the algorithm outperform the judges? Figure J.1 repeats Figure 3 within each group. Judges are inside the algorithm's frontier at their release rates for 85.4 percent of judges among Black defendants, 85.2 percent among white defendants, 88.1 percent among

male defendants and 51.2 percent among female defendants, against 87.1 percent for all defendants.

Do judges do worse than random release? The shares of judges below the random-release line are 44.7, 46.6, 52.8 and 27.0 percent, against 47.4 percent for all defendants. These shares, like those in the previous paragraph, are shares of point estimates. Sampling error spreads point estimates more than the true rates, and more so in smaller groups (Appendix G), so the shares should not be compared across groups. We have not tested whether the differences across groups are statistically significant.

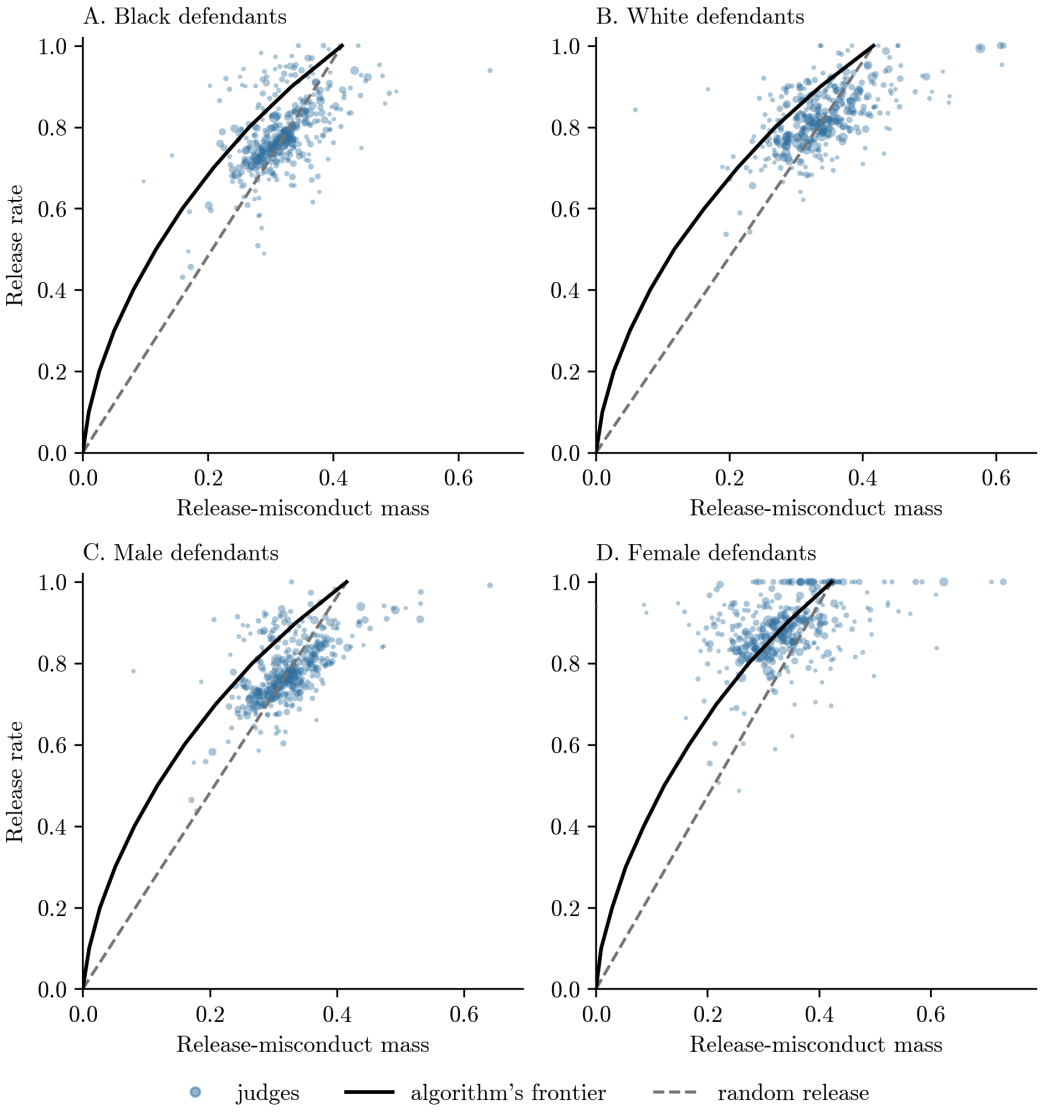
J.3 Private Information

Do judges have private information? Figure J.2 repeats Figure 6 within each group, with the test of Section 5.3.1 unchanged. Controlling the false discovery rate at 5 percent and pooling across deciles, the share of judges who reject the null of no private information is 37.6 percent for Black defendants, 53.7 percent for white defendants, 26.4 percent for male defendants and 68.2 percent for female defendants, against 25.1 percent for all defendants. Controlling the family-wise error rate at 5 percent with Holm's procedure, the shares are 19.6, 35.4, 16.7 and 51.8 percent, against 17.0 percent.

Is the evidence driven by small cells? The differences across groups mostly reflect cell sizes rather than differences in judges' information. The test rejects most often in small cells, often cells in which no released defendant commits misconduct, and the smaller the group the more such cells there are: the median rejected cell has fewer than 11 cases for female defendants and 15 for Black defendants, against 20 for all defendants, and the median cell that does not reject has 72 cases in the full sample. Among judges with at least 20 cases in every decile, the false discovery rate shares are 16.4 percent of 171 judges, 30.0 percent of 130, 7.8 percent of 205 and 10.0 percent of 30, and the Holm shares 10.5, 17.7, 3.4 and 10.0 percent, against 5.4 and 2.7 percent of 223 judges for all defendants.

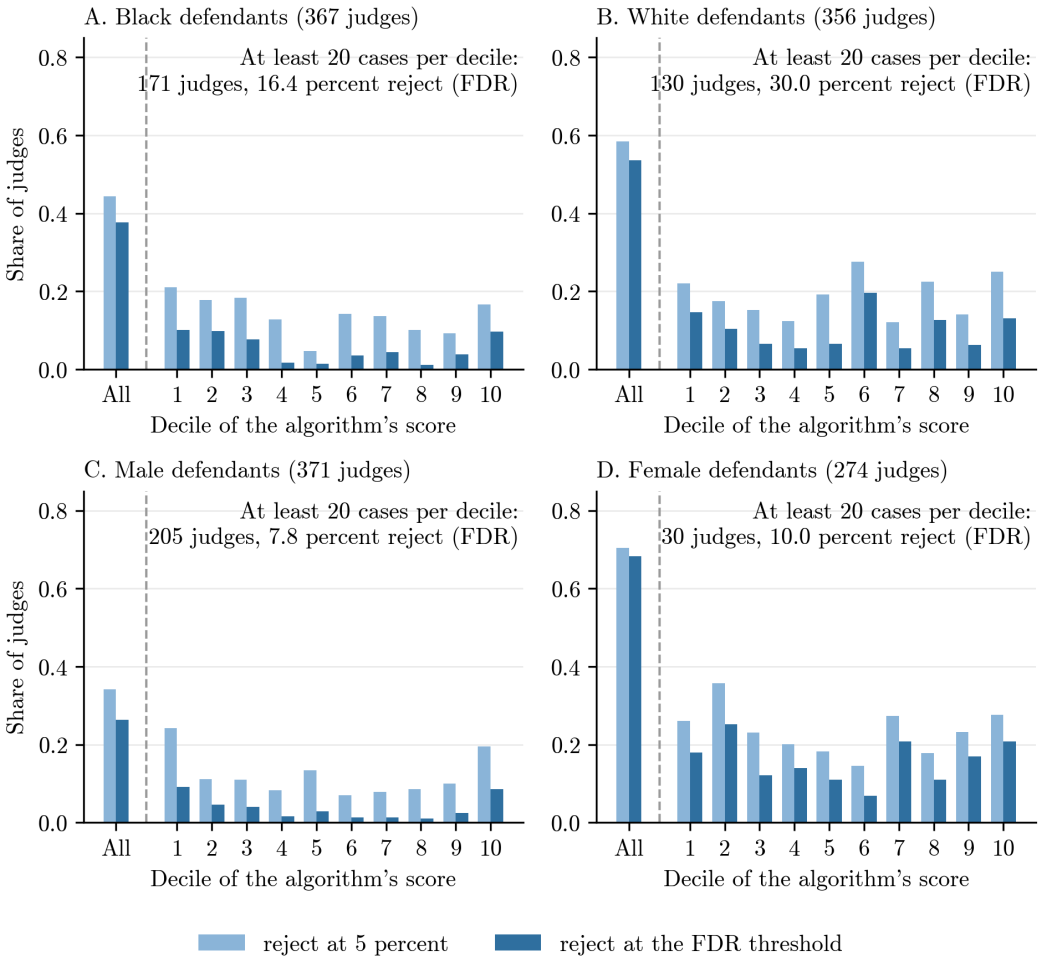
Are the judges calibrated? The test of conditional calibration rejects at the false discovery rate threshold for 12.8, 27.5, 9.4 and 42.3 percent of judges, against 7.5 percent for all defendants.

Figure J.1: The Common Comparison, by Group



Notes. As Figure 3, within each group of defendants: Black (Panel A), white (Panel B), male (Panel C) and female (Panel D). The outcome is a failure to appear or any rearrest. Judges' strata-adjusted rates (points, sized by the judge's cases in the group), the algorithm's frontier (solid line, extrapolated as in Appendix F.2) and random release (dashed line) are all estimated on the group's cases alone, with the algorithm and score deciles of the full sample. Section J.2 discusses the figure.

Figure J.2: Share of Judges With Evidence of Private Information, by Group



Notes. As Figure 6, within each group, among judges with at least 50 cases in the group. The outcome is a failure to appear or any rearrest, with decile risks extrapolated within the group as in Appendix F.3. Light bars: share of judges rejecting at 5 percent; dark bars: after the false discovery rate adjustment. The left pair pools across deciles; the others are within each decile. The text in each panel reports the share rejecting at the false discovery rate threshold among judges with at least 20 of the group's cases in every decile, with the adjustment recomputed among them. Bayesian bootstrap with 500 replications. Section J.3 discusses the figure.

Table J.1: Main Results, by Group

	All	Black	White	Male	Female
Cases	382,635	235,465	147,170	322,290	59,658
Release rate	79.6	77.8	82.4	77.8	88.9
Outcome rate among released defendants	40.8	40.5	41.4	41.6	37.4
Extrapolated risk, linear fit	42.6	40.3	44.2	42.8	38.8
Extrapolated risk, local linear fit	43.7	39.1	46.8	43.1	39.7
Extrapolated risk at the matched cutoff	33.5	30.6	36.3	33.6	34.7
Judges inside the algorithm's frontier (raw)	87.1	85.4	85.2	88.1	51.2
Judges below random release (raw)	47.4	44.7	46.6	52.8	27.0
Judges tested for private information	371	367	356	371	274
Reject no private information (FDR)	25.1	37.6	53.7	26.4	68.2
Reject no private information (Holm)	17.0	19.6	35.4	16.7	51.8
Judges with at least 20 cases in every decile	223	171	130	205	30
Of these, reject no private information (FDR)	5.4	16.4	30.0	7.8	10.0
Of these, reject no private information (Holm)	2.7	10.5	17.7	3.4	10.0
Reject calibration (FDR)	7.5	12.8	27.5	9.4	42.3

Notes. Percent, except counts of cases and judges. Each column repeats the analysis on the cases of one group, with the algorithm and score deciles of the full sample. Extrapolated risks are the fitted values at full release of Appendix F.1. Shares of judges inside the frontier and below random release are shares of point estimates. The private information and calibration tests are those of Section 5.3.1; FDR rows control the false discovery rate at 5 percent and Holm rows the family-wise error rate at 5 percent, across the judges tested in the group. Section J discusses the table.