IDENTIFYING PREDICTION MISTAKES IN OBSERVATIONAL DATA*

Ashesh Rambachan†

May 1, 2024

Abstract

Decision makers, such as doctors, judges, and managers, make consequential choices based on predictions of unknown outcomes. Do these decision makers make systematic prediction mistakes based on the available information? If so, in what ways are their predictions systematically biased? In this paper, I characterize conditions under which systematic prediction mistakes can be identified in empirical settings such as hiring, medical diagnosis, and pretrial release. I derive a statistical test for whether the decision maker makes systematic prediction mistakes under these assumptions and provide methods for estimating the ways in which the decision maker's predictions are systematically biased. I analyze the pretrial release decisions of judges in New York City, estimating that at least 20% of judges make systematic prediction mistakes about misconduct risk given defendant characteristics. Motivated by this analysis, I estimate the effects of replacing judges with algorithmic decision rules, and find that replacing judges with algorithms where systematic prediction mistakes occur dominates the status quo.

JEL Codes: C10, C55, D81, D84.

Keywords: prediction, expected utility, mistake, pretrial, algorithm.

*First version: June 2021. I thank Isaiah Andrews, Nano Barahona, Laura Blattner, Iavor Bojinov, Raj Chetty, Bo Cowgill, Will Dobbie, Xavier Gabaix, Matthew Gentzkow, Ed Glaeser, Yannai Gonczarowski, Peter Hull, Larry Katz, Jens Ludwig, Daniel Martin, Sendhil Mullainathan, Maria Polyakova, Jonathan Roth, Joshua Schwartzstein, Jesse Shapiro, Neil Shephard, Elie Tamer, and seminar participants at Berkeley, Brown, Caltech, Carnegie Mellon, Columbia, Duke, Harvard Business School, Microsoft Research, Michigan State, MIT, Northwestern Kellogg, Pennsylvania State, Princeton, Rutgers, Stanford, University of Pennsylvania, University of Toronto, Yale, the ACM SIGecom Winter Meeting, the Chamberlain Online Seminar, the NBER Economics of AI Conference, and the Virtual Digital Economy Seminar for useful comments. I also thank Hye Chang, Nicole Gillespie, Hays Golden, and Ellen Louise Dunn for assistance at the University of Chicago Crime Lab. All empirical results based on New York City pretrial data were originally reported in a University of Chicago Crime Lab technical report (Rambachan and Ludwig, 2021). I acknowledge financial support from the NSF Graduate Research Fellowship (Grant DGE1745303). All errors are my own.

†Department of Economics, MIT. 50 Memorial Drive, Cambridge, MA 02142: asheshr@mit.edu

1 Introduction

Decision makers, such as doctors, judges, and managers, make consequential decisions based on predictions of unknown outcomes. For example, in deciding whether to detain a defendant awaiting trial, a judge predicts the defendant's behavior upon release based on information such as the defendant's current criminal charge and prior arrest record. Are these decision makers making systematic prediction mistakes based on this available information? If so, which decision makers? On which decisions? And in what ways are their predictions systematically biased?

These foundational questions have renewed policy relevance and empirical life as machine learning based models increasingly replace or inform decision makers in criminal justice, health care, labor markets, and consumer finance.[1] The hope is that these tools can improve decisions by predicting more accurately than existing decision makers. In assessing whether such machine learning based models can improve decision-making, researchers therefore attempt to evaluate decision makers' implicit predictions through comparisons of their choices against those made by predictive models.[2]

Yet uncovering systematic prediction mistakes in these empirical settings is challenging as both the decision maker's preferences and information set are unknown to us. For example, we do not know how judges assess the cost of pretrial detention. Judges may uncover useful information through their courtroom interactions with defendants, but we do not observe these interactions. The decision maker's choices may therefore differ from the predictive model not because she is making systematic prediction mistakes, but rather she

[1] Risk assessment tools are used in criminal justice systems throughout the United States (Stevenson, 2018; Dobbie and Yang, 2019, 2021). Clinical risk assessments aid doctors in diagnostic and treatment decisions (Obermeyer and Emanuel, 2016; Beaulieu-Jones et al., 2019). For applications in consumer finance, see for example Einav, Jenkins and Levin (2013), Blattner and Nelson (2021), and Fuster et al. (2022). For discussions of workforce analytics and resume screening software, see Hoffman, Kahn and Li (2018), Li, Raymond and Bergman (2020), and Raghavan et al. (2020).

[2] See, for example, Hoffman, Kahn and Li (2018), Kleinberg et al. (2018), Erel et al. (2019), Li, Raymond and Bergman (2020), Jung et al. (2020), and Mullainathan and Obermeyer (2022). Comparing a decision maker's choices against a predictive model has a long tradition in psychology (e.g., Dawes, 1971, 1979; Dawes, Faust and Meehl, 1989; Camerer and Johnson, 1997; Grove et al., 2000; Kuncel et al., 2013).

has preferences that differ from the model's objective function or observes information that is unavailable to the model. While existing empirical research recognizes these challenges (e.g., Kleinberg et al., 2018; Mullainathan and Obermeyer, 2022), it lacks a unifying econometric framework for analyzing a decision maker's choices under weak assumptions about their preferences and information sets.

This paper develops an econometric framework for analyzing whether a decision maker makes systematic prediction mistakes and to characterize how their predictions are systematically biased. I clarify what can be identified about systematic prediction mistakes from data and empirically relevant assumptions about behavior, and map those assumptions into statistical inferences about systematic prediction mistakes. This framework, therefore, robustly measures key behavioral mechanisms underlying the tradeoffs between human and algorithmic decision making.

I consider empirical settings, such as pretrial release, medical diagnosis, and hiring, in which a decision maker makes choices for many individuals based on a prediction of some unknown outcome using each individual's characteristics. The characteristics are observable to both the decision maker and the researcher. There is a *missing data* problem: the researcher only observes the outcome conditional on the decision maker's choices (e.g., we only observe a defendant's behavior upon release if a judge released them).

In these empirical settings, I explore the nonparametric restrictions imposed on the decision maker's choices by expected utility maximization, which models the decision maker as maximizing some (unknown to us) utility function at some beliefs about the outcome given the characteristics as well as any distribution of private information. Due to the missing data problem, the true conditional distribution of the outcome given the characteristics is partially identified. The expected utility maximization model therefore only restricts the decision maker's beliefs to lie somewhere in this identified set, a restriction I call "accurate beliefs." If there exists no utility function, accurate beliefs, nor any private information that rationalizes their observed choices, I say the decision maker is making "systematic prediction

mistakes" based on the characteristics of individuals.

I provide an empirical characterization of expected utility maximization at accurate beliefs over an economically rich class of utility functions. This characterization implies systematic prediction mistakes are *untestable* without further assumptions. If either all observed characteristics of individuals directly affect the decision maker's utility function or the missing data can take any value, then the decision maker's choices can always be rationalized. However, placing an exclusion restriction on which characteristics may directly affect the decision maker's utility function and informative bounds on the missing data restores the testability of expected utility maximization behavior. Variation in choices across characteristics that do not directly affect the utility function must only arise due to variation in beliefs and their beliefs must be Bayes-plausible with respect to some conditional distribution of the outcome given the characteristics in the identified set, together implying testable restrictions.

With this framework in place, I further establish that the data are informative about the magnitudes of the decision maker's systematic prediction mistakes. As a computational device, I extend the behavioral model to only require that the decision maker's choices approximately maximize expected utility, meaning that their choices must only be within some expected utility cost of being optimal without taking a stand on what drives the decision maker's misoptimizations. I derive bounds on the total expected utility cost to the decision maker of their systematic prediction mistakes and the share of systematic prediction mistakes in their decisions.

I then analyze whether the data are informative about the ways in which the decision maker's beliefs are systematically biased. I allow the decision maker to have possibly inaccurate beliefs about the outcome, no longer requiring their beliefs to lie in the identified set for the true conditional distribution of the outcome given the characteristics. This takes no stand on the behavioral foundations for the decision maker's inaccurate beliefs, and so it encompasses various mechanisms like inattention to characteristics, representativeness, or

salience (e.g., Handel and Schwartzstein, 2018; Bordalo et al., 2016; Bordalo, Gennaioli and Shleifer, 2021). For a binary outcome, I bound a parameter that summarizes the extent to which the decision maker's beliefs overreact or underreact to the characteristics of individuals. These bounds may indicate whether the decision maker's beliefs vary more ("overreact") or less than ("underreact") the true conditional distribution of the outcome across values of the characteristics.

As an empirical illustration, I analyze the pretrial system in New York City, in which judges decide whether to release defendants awaiting trial based on a prediction of whether they will fail to appear in court.[3] For each judge, I observe the conditional probability that she releases a defendant given a rich set of characteristics (e.g., race, age, current charge, prior criminal record, etc.) as well as the conditional probability that a released defendant fails to appear in court. The conditional failure to appear rate among detained defendants is unobserved due to the missing data problem.

If all defendant characteristics may directly affect the judge's utility function or the conditional failure to appear rate among defendants detained by this judge may take any value, then my identification results establish that the judge's release decisions are always consistent with expected utility maximization behavior at accurate beliefs. Absent further assumptions, the judge's release decisions could reflect either a utility function that varies richly based on defendant characteristics or private information. However, empirical researchers may be willing to assume, for example, that while judges may engage in taste-based discrimination on a defendant's race, other defendant characteristics such as prior pretrial misconduct history only affect judges' beliefs about failure to appear risk. Judges in New York City are quasi-randomly assigned to defendants, which implies bounds on the conditional failure to appear rate among detained defendants. Given such utility exclusion restrictions and quasi-experimental bounds on the missing data, my identification results establish that expected

[3]In the New York City pretrial system, judges decide whether to release defendants prior to their trial without conditions ("on own recognizance") or set monetary bail conditions. In Section 5.4.2, I report the robustness of my findings to alternative definitions of the pretrial decision.

utility maximization behavior at accurate beliefs is falsified by *misrankings* in the judge's release decisions. Holding fixed defendant characteristics that may directly affect utility (e.g., among defendants of the same race), do all released defendants have a lower failure to appear rate than the upper bound on the failure to appear rate of all defendants detained by this judge? If not, there exists no utility function satisfying the conjectured exclusion restriction nor private information such that the judge's choices maximize expected utility at accurate beliefs about failure to appear risk given defendant characteristics.

By testing for such misrankings in the pretrial release decisions of individual judges, I estimate, as a lower bound, that at least 20% of judges in New York City from 2008-2013 make systematic prediction mistakes about failure to appear risk based on defendant characteristics. Under alternative utility exclusion restrictions, there exists no utility function nor distribution of private information such that the release decisions of these judges would maximize expected utility at accurate beliefs about failure to appear risk. These systematic prediction mistakes occur over many defendants, are costly to judges in an expected utility sense, and arise because judges' beliefs underreact to variation in failure to appear risk based on defendant characteristics between predictably low risk and predictably high risk defendants. Rejections of expected utility maximization behavior at accurate beliefs are driven by release decisions on defendants at the tails of the predicted risk distribution. These findings are robust to alternative definitions of the pretrial misconduct outcome, alternative definitions of the pretrial decision, and alternative empirical strategies for bounding the failure to appear rate among detained defendants.

Finally, I analyze the effects of replacing particular judges with algorithmic decision rules in the New York City pretrial system. The policymaker's tradeoff between a judge's decisions and an algorithmic decision rule depend on whether the judge makes systematic prediction mistakes about failure to appear risk and if so on which defendants, whether the judge is misaligned and optimizing a different objective than the policymaker, and finally whether the judge observes any useful private information that is unavailable to the algorithm. By

allowing for these three competing forces, the preceding behavioral analysis informs our understanding of the possible tradeoff between human and algorithmic decision making.

I estimate the effects of replacing judges who were found to make systematic prediction mistakes with an algorithmic decision rule. Replacing these judges with an algorithmic decision rule only where systematic prediction mistakes occur at the tails of the predicted risk distribution weakly dominates the status quo, and can lead to up to 20% improvements in worst-case expected social welfare (measured as a weighted average of the failure to appear rate among released defendants and the pretrial detention rate). Since there exists no utility function nor distribution of private information at which their decisions over these defendants maximize expected utility, replacing judges with an algorithmic decision rule over the tails of the predicted risk distribution can be a free lunch by correcting systematic prediction mistakes. Fully replacing judges with an algorithmic decision rule, by contrast, has ambiguous effects that depend on the policymaker's objective due to a tradeoff between the extent of misalignment and the value of private information.

In related work, Kleinberg et al. (2018) directly compare the pretrial release decisions of all judges in New York City against an estimated, machine learning based decision rule. Mullainathan and Obermeyer (2022) analogously compare doctors' stress testing decisions against a machine learning based decision rule. Viewed through the lens of my identification analysis, Kleinberg et al. (2018) is limited to making statements about judge decision making under several assumptions: first, that judges' utility functions do not vary based on defendant characteristics; second, utility functions do not vary across judges; and third, that private information does not vary across judges. I conduct my analysis judge-by-judge, allowing each judge's utility function to flexibly vary based on defendant characteristics and heterogeneity in both utility functions and private information across judges.

Analyzing radiologists, Chan, Gentzkow and Yu (2022) estimate a structural model in which decision makers' beliefs are summarized by a normally distributed signal structure (see also Abaluck et al., 2016; Arnold, Dobbie and Hull, 2022). I nonparametrically model the

decision maker's beliefs as arising from observable characteristics and private information, allowing me to separately explore these two components of the information environment. In exchange for this flexibility, testing whether choices are consistent with expected utility maximization at accurate beliefs requires an exclusion restriction on which characteristics affect the decision maker's utility function and any informative bounds on the missing data, although I allow the utility function to vary arbitrarily across non-excluded characteristics.

Finally, I build on a growing literature in microeconomic theory that derives the testable implications of behavioral models in state-dependent stochastic choice (SDSC) data (e.g., Caplin and Martin, 2015; Caplin and Dean, 2015; Caplin et al., 2020). While useful in analyzing lab-based experiments, such results have had limited applicability due to the difficulty of collecting SDSC data (Gabaix, 2019; Caplin, 2021). I focus on common empirical settings in which the data suffer from a missing data problem, showing that these settings can approximate ideal SDSC data by using quasi-experimental variation to address the missing data problem. Martin and Marx (2022) study the identification of taste-based discrimination in a binary choice experiment. The setting I consider nests theirs by allowing for several key features of observational data such as missing data, multi-valued outcomes, and multiple choices. I follow in spirit of the information design literature (e.g., Bergemann and Morris, 2016, 2019) by asking whether there exists *any* private information that could rationalize the decision maker's choices. Recent papers take this approach in different settings or to answer different questions, such as Syrgkanis, Tamer and Ziani (2018) in auctions, Magnolfi and Roncoroni (2021) in entry games, Bergemann, Brooks and Morris (2022) on the welfare effects of unknown information structures, and Gualdani and Sinha (2020) in discrete choice models.

2 Expected utility maximization at accurate beliefs

A decision maker makes choices for many individuals based on the prediction of an unknown outcome using each individual's characteristics. Under what conditions do the

decision maker's choices maximize expected utility at some utility function, accurate beliefs given the characteristics, and any additional private information?

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 an unknown outcome $y^* \in \mathcal{Y}$. The random vector $(X, C, Y^*) \sim P(\cdot)$ defined over $\mathcal{X} \times \{0,1\} \times \mathcal{Y}$ summarizes the joint distribution of the characteristics, the decision maker's choices, and outcomes over all individuals. I assume the characteristics and outcome have finite support, and there exists $\delta > 0$ such that $P(x) := P(X = x) \geq \delta$ for all $x \in \mathcal{X}$.

We observe the characteristics of each individual as well as the decision maker's choice. There is a *missing data* or *selective labels* problem: we only observe Y^* if the decision maker selected $C = 1$ (Rubin, 1976; Kleinberg et al., 2018). Defining $Y := C \cdot Y^*$, the *observable data* is the joint distribution $(X, C, Y) \sim P(\cdot)$. I assume this joint distribution is known to focus on the identification challenges in this setting. The decision maker's conditional choice probabilities are $\pi_c(x) := P(C = c \mid X = x)$ for $c \in \{0,1\}$ and $x \in \mathcal{X}$, and the observable conditional outcome probabilities are $P_1(y^* \mid x) := P(Y^* = y^* \mid C = 1, X = x)$ for $x \in \mathcal{X}$. The conditional outcome probabilities $P_0(y^* \mid x) := P(Y^* = y^* \mid C = 0, X = x)$ are not identified due to the missing data problem. The true outcome probabilities $P(y^* \mid x) := P(Y^* = y^* \mid X = x)$ are also not identified as a consequence.

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

Example (Pretrial Release)**.** A judge decides whether to detain or release defendants $C \in \{0,1\}$ awaiting trial (e.g., Arnold, Dobbie and Yang, 2018; Kleinberg et al., 2018; Arnold, Dobbie and Hull, 2022). The outcome $Y^* \in \{0,1\}$ is whether a defendant would fail to appear in court if released. The characteristics X summarize recorded information about the defendant such as demographics, the defendant's current charges, and the defendant's prior arrest/conviction record. The judge's conditional release rate $\pi_1(x)$ and the conditional

failure to appear rate among released defendants $P_1(y^* \mid x)$ are identified. The conditional failure to appear rate among detained defendants $P_0(y^* \mid x)$ is not identified. ▲

Example (Medical Testing and Diagnosis)**.** A doctor decides whether to conduct a medical test or make a particular diagnosis (e.g., Abaluck et al., 2016; Chan, Gentzkow and Yu, 2022). For example, shortly after an emergency room visit, a doctor decides whether to conduct a stress test on patients $C \in \{0, 1\}$ to determine whether they had a heart attack (Mullainathan and Obermeyer, 2022). The outcome $Y^* \in \{0, 1\}$ is whether the patient had a heart attack. The characteristics X summarize recorded information about the patient such as demographics, reported symptoms, and prior medical history. The doctor's conditional stress testing rate $\pi_1(x)$ and the conditional heart attack rate among stress tested patients $P_1(y^* \mid x)$ are identified. The conditional heart attack rate among untested patients $P_0(y^* \mid x)$ is not identified. ▲

Example (Hiring)**.** A hiring manager decides whether to hire job applicants $C \in \{0, 1\}$ (Autor and Scarborough, 2008; Hoffman, Kahn and Li, 2018; Frankel, 2021). The outcome Y^* is a vector of on-the-job productivity measures, such as length of tenure since turnover may be costly. The characteristics X are recorded information about the applicant such as demographics, education level, and prior work history. The manager's conditional hiring rate $\pi_1(x)$ and the conditional distribution of on-the-job productivity measures among hired applicants $P_1(y^* \mid x)$ are identified. The distribution of on-the-job productivity measures among rejected applicants $P_0(y^* \mid x)$ is not identified. ▲

Such screening decisions are a leading class of "prediction policy problems" (Kleinberg et al., 2015). Other examples include job interviews (e.g., Li, Raymond and Bergman, 2020), loan approvals (e.g., Fuster et al., 2022), and child welfare screenings (Chouldechova et al., 2018).

In the main text, I make two simplifying assumptions: (i) the decision maker only faces two choices; and (ii) the decision maker's choice does not have a direct causal effect on the outcome. Online Appendix B generalizes my identification results to treatment assignment

problems, in which the decision maker allocates individuals to alternative treatments that causally affect the outcome, nesting the main text as a special case.

As notation, let $\Delta(\mathcal{A})$ denote the set of all probability distributions on a finite set $\mathcal{A}$. For $c \in \{0,1\}$, let $P_c(\cdot \mid x) \in \Delta(\mathcal{Y})$ denote the vector of conditional outcome probabilities given choice $C = c$ and characteristics $X = x$. Let $P(\cdot \mid x) \in \Delta(\mathcal{Y})$ denote the vector of true outcome probabilities given characteristics $X = x$.

2.2 Bounds on the missing data

I model the researcher's assumptions about the missing data problem in the form of researcher-specified bounds on the unobserved conditional outcome probabilities.

Assumption 2.1. For each $x \in \mathcal{X}$, there exists a known subset $\mathcal{B}_x \subseteq \Delta(\mathcal{Y})$ satisfying $P_0(\cdot \mid x) \in \mathcal{B}_x$.

By modelling the researcher's assumptions in terms of generic bounds $\mathcal{B}_x$, Assumption 2.1 captures many empirical strategies. In some cases, researchers may wish to analyze the decision maker's choices without placing any further assumptions on the missing data, which corresponds to $\mathcal{B}_x = \Delta(\mathcal{Y})$. Researchers may instead use quasi-experimental variation or introduce additional restrictions to bound the unknown conditional outcome probabilities, in which case $\mathcal{B}_x$ may depend on identified features of the data as I discuss in Section 3.

Under Assumption 2.1, the identified set for the true outcome probabilities given $x \in \mathcal{X}$, denoted $\mathcal{H}(P(\cdot \mid x); \mathcal{B}_x)$, equals the set of $\widetilde{P}(\cdot \mid x) \in \Delta(\mathcal{Y})$ satisfying $\widetilde{P}(y^* \mid x) = \widetilde{P}_0(y^* \mid x)\pi_0(x) + P_1(y^* \mid x)\pi_1(x)$ for all $y^* \in \mathcal{Y}$ and some $\widetilde{P}_0(\cdot \mid x) \in \mathcal{B}_x$.

2.3 Behavioral model

I examine the restrictions placed on the decision maker's choices by expected utility maximization at accurate beliefs. Under this model, the decision maker maximizes expected utility given an information set for each individual that consists of their characteristics and some additional private information which I denote by the random variable V. For example, doctors may learn useful information about the patient's current health in an exam, and

judges may interact with defendants during the pretrial release hearing; but these interactions may not be recorded.

Suppose the researcher partitions the characteristics $x := (x_I, x_E)$ with $\mathcal{X} = \mathcal{X}_I \times \mathcal{X}_E$. The expected utility maximization model is summarized by a utility function and a joint distribution over the characteristics, private information, choices and outcomes, denoted $(X, V, C, Y^*) \sim Q(\cdot)$, that satisfies three conditions.

Definition 2.1. A *utility function* $u\colon \{0,1\} \times \mathcal{Y} \times \mathcal{X}_I \to \mathbb{R}$ specifies the payoff associated with each choice-outcome pair at characteristics $x_I \in \mathcal{X}_I$. Let $\mathcal{U}$ denote the feasible set of utility functions specified by the researcher.

Definition 2.2. The decision maker's choices are *consistent with expected utility maximization* if there exists a utility function $u \in \mathcal{U}$ and joint distribution $(X, V, C, Y^*) \sim Q(\cdot)$ satisfying

i. Expected Utility Maximization: For all $c \in \{0,1\}$, $c' \neq c$, $(x, v) \in \mathcal{X} \times \mathcal{V}$ such that $Q(c \mid x, v) > 0$,

$$\mathbb{E}_Q\left[u(c, Y^*; X_I) \mid X = x, V = v\right] \geq \mathbb{E}_Q\left[u(c', Y^*; X_I) \mid X = x, V = v\right].$$

ii. Information Set: $C \perp\!\!\!\perp Y^* \mid X, V$ under $Q(\cdot)$.

iii. Data Consistency: For all $x \in \mathcal{X}$, there exists $\widetilde{P}_0(\,\cdot \mid x) \in \mathcal{B}_x$ satisfying

$$Q(x, c, y^*) = \begin{cases} P_1(y^* \mid x)\pi_1(x)P(x) \text{ if } c = 1 \\ \widetilde{P}_0(y^* \mid x)\pi_0(x)P(x) \text{ if } c = 0 \end{cases}$$

for all $y^* \in \mathcal{Y}$.

2.3.1 Interpreting the utility exclusion restriction

The key behavioral assumption is an *exclusion restriction* on the decision maker's utility

function. Only the characteristics X_I are included in the decision maker's utility function. The remaining characteristics X_E are excluded from the decision maker's utility function and only affect the decision maker's beliefs.

In many settings, researchers either already make such utility exclusion restrictions or reason about their plausibility. In medical testing and diagnosis, researchers often assume that a doctor's payoffs are constant across patients or only depend on a limited set of characteristics (e.g., Abaluck et al., 2016; Chan, Gentzkow and Yu, 2022; Mullainathan and Obermeyer, 2022). Structural analyses of pretrial decisions assume that a judge's payoffs may only directly depend on the race of a defendant (e.g., Arnold, Dobbie and Hull, 2022), and recent work considers the validity of marginal outcome tests under alternative utility exclusion restrictions (Becker, 1957; Arnold, Dobbie and Yang, 2018; Canay, Mogstad and Mountjoy, 2020). Other researchers empirically explore whether judges may be more lenient towards younger defendants (Stevenson and Doleac, 2022) or more harsh towards defendants charged with violent crimes (Kleinberg et al., 2018).

Since this is a substantive economic assumption, I discuss two ways researchers may specify such utility exclusion restrictions. First, the researcher may conduct a sensitivity analysis, reporting how their conclusions vary as the choice of utility exclusion restriction varies. Such a sensitivity analysis summarizes how flexible the decision maker's utility function must be across characteristics to rationalize choices and how behavioral conclusions vary across plausible assumptions. Second, the exclusion restriction may also be normatively motivated, summarizing social or legal restrictions on what characteristics ought not to directly enter the decision maker's utility function.

2.3.2 Accurate beliefs and systematic prediction mistakes

If Definition 2.2 is satisfied, then the decision maker's implied beliefs about the outcome given the characteristics, denoted $Q(\cdot \mid x) \in \Delta(\mathcal{Y})$, lie in the identified set for the true outcome probability $P(\cdot \mid x)$. That is, $Q(\cdot \mid x) \in \mathcal{H}(P(\cdot \mid x); \mathcal{B}_x)$ for all $x \in \mathcal{X}$ as a consequence of Data Consistency, and the decision maker's implied beliefs $Q(\cdot \mid x)$ are

accurate in this weak sense if their choices are consistent with expected utility maximization.

Put another way, if the decision maker's choices are inconsistent with expected utility maximization, then there exists no utility function nor private information such that their choices would maximize expected utility at *any* beliefs in the identified set for the true outcome probability $P(\cdot \mid x)$. The decision maker is behaving as-if their implied beliefs given the characteristics are systematically mistaken in this strong sense.

Definition 2.3. The decision maker is making *systematic prediction mistakes* over the class of utility functions $\mathcal{U}$ if their choices are inconsistent with expected utility maximization.

While the behavioral model restricts the decision maker's implied beliefs to lie in the identified set for the true outcome probability, it is otherwise agnostic about how the decision maker arrives at those beliefs. In this sense, if the decision maker is making systematic prediction mistakes, then their choice behavior is inconsistent with *any* model of belief formation that leads to beliefs in the identified set for the true outcome probabilities.

Furthermore, the interpretation of a systematic prediction mistake is tied to both the researcher-specified bounds on the missing data (Assumption 2.1) and the feasible set of utility functions $\mathcal{U}$ (Definition 2.1). Less informative bounds on the missing data imply there are more candidate values of the missing conditional outcome probabilities and, in turn, more candidate values of the true outcome probabilities that may rationalize choices. A larger feasible set of utility functions $\mathcal{U}$ analogously implies that expected utility maximization places fewer restrictions on behavior as the researcher is entertaining a larger set of utility functions that may rationalize choices. Definition 2.3 must therefore be interpreted in the context of the researcher's assumptions on both the missing data and the decision maker's utility function.

3 Identifying systematic prediction mistakes in screening decisions

I characterize conditions under which expected utility maximization at accurate beliefs has testable implications. Testing whether the decision maker's choices are consistent with expected utility maximization at accurate beliefs is equivalent to testing moment inequalities under these conditions.

3.1 Characterization result

I characterize when the decision maker's choices are consistent with expected utility maximization over the class of *linear* utility functions.

Assumption 3.1 (Vector-valued outcome)**.** For $K \geq 1$, the outcome satisfies $y^* := (y_1^*, \ldots, y_K^*) \in [0,1]^K$.

Definition 3.1. Under Assumption 3.1, the class of *linear utility functions* is the set of utility functions satisfying $u(c, y^*; x_I) = \sum_{k=1}^{K} u_{1,k}(x_I) y_k^* c + u_{0,k}(x_I)(1 - y_k^*)(1 - c)$, where $u_{1,k}(x_I), u_{0,k}(x_I) \leq 0$, $|u_{1,k}(x_I) + u_{0,k}(x_I)| = 1$ for all $x_I \in \mathcal{X}_I$.

This is an economically rich class that captures many common empirical intuitions. The parameters $u_{1,k}(x_I), u_{0,k}(x_I) \leq 0$ summarize the costs of ex-post errors for each outcome (i.e., selecting $C = 1$ when Y_k^* is large and selecting $C = 0$ when Y_k^* is small respectively). It places no restrictions on how costs vary across included characteristics X_I and outcomes Y_k^*. In the pretrial release example, defining $Y^* = Y_1^* \in \{0,1\}$ to be whether a defendant would fail to appear in court, the class of linear utility functions assumes it is costly for the judge to detain a defendant that would not fail to appear or release a defendant that would fail to appear, but places no restrictions on how these costs vary across included defendant characteristics X_I, such as defendant race, age, or charge severity. If instead $Y^* = (Y_1^*, Y_2^*)$ is whether a defendant would fail to appear in court $Y_1^* \in \{0,1\}$ and be re-arrested $Y_2^* \in \{0,1\}$, the class of linear utility functions also places no restriction on the relative cost of releasing

a defendant that would fail to appear versus be re-arrested.

For $x_I \in \mathcal{X}_I$ and $c \in \{0,1\}$, define $\Pi_c(x_I) := \{x_E \in \mathcal{X}_E \colon \pi_c(x_I, x_E) > 0\}$. Let $\bar{Y}^* := \sum_{k=1}^K Y_k^*$, $\mu_c(x) := \mathbb{E}[\bar{Y}^* \mid C = c, X = x]$ for $c \in \{0,1\}$, and $\bar{\mu}_0(x) := \max_{\tilde{P}_0(\cdot|x) \in \mathcal{B}_x} \mu_0(x)$.

Theorem 3.1. *Suppose Assumption 3.1 holds. The decision maker's choices are consistent with expected utility maximization at some linear utility function if and only if, for all $x_I \in \mathcal{X}_I$,*

$$\max_{x_E \in \Pi_1(x_I)} \mu_1(x_I, x_E) \leq \min_{x_E \in \Pi_0(x_I)} \bar{\mu}_0(x_I, x_E) \tag{1}$$

Otherwise, the decision maker is making systematic prediction mistakes over the class of linear utility functions.

Over the class of linear utility functions, expected utility maximization requires the decision maker to make choices according to an incomplete threshold rule based on the expectation for $\bar{Y}^*$ under their beliefs given the characteristics and their private information. The threshold may vary across included characteristics X_I, and it is incomplete since the behavioral model takes no stand on how possible indifferences are resolved. The proof of Theorem 3.1 shows that the conditional outcome probabilities summarize all possible beliefs that could arise. The inequalities (1) check whether any value of the conditional outcome probabilities consistent with the researcher's bounds (Assumption 2.1) could reproduce the decision maker's choices under such a threshold rule.[4]

3.2 When are systematic prediction mistakes identifiable?

By examining when the inequalities in Theorem 3.1 are always satisfied, I characterize leading cases in which we cannot identify systematic prediction mistakes.

Corollary 3.1. *Suppose Assumption 3.1 holds. The decision maker's choices are consistent with expected utility maximization at accurate beliefs and some linear utility function if either:*

[4]Theorem 3.1 builds on the "no-improving action switches" inequalities, which were originally derived by Caplin and Martin (2015) to analyze choice behavior in state-dependent stochastic choice data from experiments.

(i) all characteristics directly affect utility (i.e, $\mathcal{X} = \mathcal{X}_I$) and $\mu_1(x_I) \leq \bar{\mu}_0(x_I)$ for all $x_I \in \mathcal{X}_I$; or (ii) $\bar{\mu}_0(x) = K$ for all $x \in \mathcal{X}$.

If all characteristics are included in the decision maker's utility function (i.e., $\mathcal{X} = \mathcal{X}_I$), then the decision maker's choices are consistent with expected utility maximization whenever the researcher's bounds are compatible with the existence of private information (Corollary 3.1(i)). More precisely, if the conditional expectation of $\bar{Y}^*$ given $C = 0$ may be at least as large as the observed conditional expectation of $\bar{Y}^*$ given $C = 1$ under the researcher's assumptions, then a threshold rule in which the threshold richly varies across the characteristics can always rationalize the decision maker's choices. If the researcher's bounds allow for the existence of private information in this weak sense, then a utility exclusion restriction is necessary to identify systematic prediction mistakes. Corollary 3.1(ii), however, establishes that such an exclusion restriction alone may be insufficient. Absent informative bounds on the unobservable conditional outcome probabilities, the decision maker's choices may always be rationalized by the extreme case in which the decision maker's private information is perfectly predictive of the unknown outcome. In this sense, identifying systematic prediction mistakes over the class of linear utility functions requires behavioral assumptions that place an exclusion restriction on the decision maker's utility function and econometric assumptions that generate informative bounds on the unobservable conditional outcome probabilities.

Under such assumptions, Theorem 3.1 provides interpretable conditions for identifying systematic prediction mistakes. Holding fixed $x_I \in \mathcal{X}_I$, does there exist some excluded characteristic $x_E \in \mathcal{X}_E$ such that the largest possible expected value of $\bar{Y}$ given $C = 0$ is strictly lower than the observed expected value of $\bar{Y}$ given $C = 1$ at some other excluded characteristic $x'_E \in \mathcal{X}_E$? If so, the decision maker could do strictly better by raising their probability of selecting choice $C = 0$ at x'_E and lowering their probability of selecting choice $C = 1$ at x_E no matter her linear utility function, implied beliefs given the characteristics, and private information.

In the pretrial release example, we may suspect the judge engages in taste-based dis-

crimination based on defendant race. Checking whether the judge's release decisions are consistent with expected utility maximization at accurate beliefs about failure to appear risk requires checking, among defendants of the same race, whether there exists some group of released defendants with a higher failure to appear rate than the worst-case failure to appear rate of some group of detained defendants among defendants. If so, the judge must be misranking defendants based on failure to appear risk given their characteristics, and their choices are inconsistent with expected utility maximization at any accurate beliefs, private information, and linear utility function that depends arbitrarily on defendant race. These *misrankings* characterize the joint null hypothesis that the decision maker's choices are consistent with expected utility maximization at accurate beliefs and some linear utility function satisfying the conjectured utility exclusion restriction.

3.3 Constructing bounds on the missing data

Suppose there is a randomly assigned instrument that generates variation in the decision maker's choice probabilities. Such instruments commonly arise, for example, through the random assignment of decision makers – judges may be randomly assigned to defendants in pretrial release (e.g., Kling, 2006; Dobbie, Goldin and Yang, 2018; Arnold, Dobbie and Yang, 2018; Kleinberg et al., 2018; Arnold, Dobbie and Hull, 2022), and doctors may be randomly assigned to patients in medical testing (e.g., Abaluck et al., 2016; Chan, Gentzkow and Yu, 2022).

Assumption 3.2 (Conditionally Random Instrument)**.** Let $Z \in \mathcal{Z}$ be a finite support instrument. The joint distribution $(X, Z, C, Y^*) \sim P(\cdot)$ satisfies $Y^* \perp\!\!\!\perp Z \mid X$, and there exists some $\delta > 0$ such that $P(x, z) := P(X = x, Z = z) \geq \delta$ for all $(x, z) \in \mathcal{X} \times \mathcal{Z}$.

The conditional expectation $\mu_0(x, z) := \mathbb{E}[\bar{Y}^* \mid C = 0, X = x, Z = z]$ is partially identified under Assumption 3.2. In the case where the instrument arises through the random assignment of decision makers, the identified set for $\mu_0(x, z)$ corresponds to sharp bounds on the conditional outcome probabilities for a single decision maker.

Proposition 3.1. *Suppose Assumptions 3.1-3.2 hold. For any* $(x, z) \in \mathcal{X} \times \mathcal{Z}$ *with* $\pi_0(x, z) > 0$*, the identified set for* $\mu_0(x, z)$ *is the interval* $[\underline{\mu}_0(x, z), \overline{\mu}_1(x, z)]$*, where*

$$\underline{\mu}_0(x, z) = \max\left\{\frac{\underline{\mu}(x) - \mu_1(x, z)\pi_1(x, z)}{\pi_0(x, z)}, 0\right\}, \text{ and } \overline{\mu}_0(x, z) = \min\left\{\frac{\overline{\mu}(x) - \mu_1(x, z)\pi_1(x, z)}{\pi_0(x, z)}, 1\right\}$$

for $\underline{\mu}(x) = \max_{\tilde{z} \in \mathcal{Z}}\{\mu_1(x, \tilde{z})\pi_1(x, \tilde{z})\}$ *and* $\overline{\mu}_0(x) = \min_{\tilde{z} \in \mathcal{Z}}\{K\pi_0(x, \tilde{z}) + \mu_1(x, \tilde{z})\pi_1(x, \tilde{z})\}$.

Online Appendix C.1 derives bounds under the assumption that the instrument is quasi-randomly assigned conditional on only some characteristic such as courtroom-by-time indicators in the pretrial release example, which I use in the empirical application to the New York City pretrial system.

Under the expected utility maximization model, Assumption 3.2 only requires that if the decision maker's choices are consistent with expected utility maximization at some utility function $u \in \mathcal{U}$ and joint distribution $(X, Z, V, C, Y^*) \sim Q$, then $Y^* \perp\!\!\!\perp Z \mid X$ under $Q(\cdot)$. Requiring that the decision maker's beliefs be accurate imposes that the instrument cannot affect their beliefs about the outcome given the characteristics. Both the utility function and private information can otherwise richly vary with the instrument. In the pretrial release example, if all judges make choices as-if they maximize expected utility at accurate beliefs and judges are randomly assigned to defendants, then all judges must have the same beliefs about failure to appear risk given defendant characteristics. Judges may still richly differ from one another in their utility functions and private information. In this sense, these bounds do not require monotonicity (e.g., see de Chaisemartin, 2017; Frandsen, Lefgren and Leslie, 2019). These bounds also require no parametric extrapolation across decision makers (e.g., Arnold, Dobbie and Hull, 2022; Angelova, Dobbie and Yang, 2023).

By setting the researcher's bounds (Assumption 2.1) to be consistent with the derived instrumental variable bounds (Proposition 3.1), I apply Theorem 3.1 to test whether the decision maker's choices are consistent with expected utility maximization at accurate beliefs and some linear utility function.

Proposition 3.2. *Suppose Assumptions 3.1-3.2 hold, and* $0 < \pi_1(x,z) < 1$ *for all* $(x,z) \in \mathcal{X} \times \mathcal{Z}$*. The decision maker's choices at* $z \in \mathcal{Z}$ *are consistent with expected utility maximization at some linear utility function if and only if, for all* $x_I \in \mathcal{X}_I$*, pairs* $x_E, \tilde{x}_E \in \mathcal{X}_E$ *and* $\tilde{z} \in \mathcal{Z}$*,*

$$\mu_1(x_I, x_E, z) - \overline{\mu}_{0,\tilde{z}}(x_I, \tilde{x}_E, z) \leq 0, \tag{2}$$

where $\overline{\mu}_{0,\tilde{z}}(x,z) = \frac{K\pi_0(x,\tilde{z}) + \mu_1(x,\tilde{z})\pi_1(x,\tilde{z})}{\pi_0(x,z)} - \frac{\mu_1(x,z)\pi_1(x,z)}{\pi_0(x,z)}$.

Equation (2) is a system of many moment inequalities. The number of moment inequalities equals $|\mathcal{X}_I| \cdot |\mathcal{X}_E|^2 \cdot (|\mathcal{Z}| - 1)$, and grows rapidly with the support of the characteristics and instruments. In empirical applications, the number of moment inequalities will typically be extremely large, posing a practical challenge as the number of observations in each cell of characteristics $x \in \mathcal{X}$ can be extremely small. I return to this problem in the empirical application to the New York City pretrial system.

4 Characterizing systematic prediction mistakes in screening decisions

Researchers can identify systematic prediction mistakes over the class of linear utility functions by searching for misrankings in the decision maker's choices. Such misrankings are further informative about the magnitudes of the decision maker's systematic prediction mistakes and the ways in which the decision maker's beliefs are systematically biased.

4.1 Bounding the costs and share of systematic prediction mistakes

To characterize the magnitudes of the decision maker's systematic prediction mistakes, I weaken Definition 2.2 to only require that the decision maker's choices *approximately* maximize expected utility.

Definition 4.1. The decision maker's choices *approximately maximize* expected utility at accurate beliefs and expected utility costs $\epsilon := \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ if there exists a utility function $u \in \mathcal{U}$ and joint distribution $(X, V, C, Y^*) \sim Q(\cdot)$ satisfying:

i. Approximate Expected Utility Maximization: For all $c \in \{0, 1\}$, $c' \neq c$, $(x, v) \in \mathcal{X} \times \mathcal{V}$ such that $Q(c \mid x, v) > 0$,

$$\mathbb{E}_Q\left[u(c, Y^*; X_I) \mid X = x, V = v\right] \geq \mathbb{E}_Q\left[u(c', Y^*; X_I) \mid X = x, V = v\right] - \epsilon(x).$$

and (ii) Information Set, (iii) Data Consistency as in Definition 2.2. The *identified set of expected utility costs* is the set of $\epsilon := \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ such that there exists $u \in \mathcal{U}$ and $(X, V, C, Y^*) \sim Q(\cdot)$ satisfying (i)-(iii).

I use Definition 4.1 to characterize the extent to which the decision maker's choices deviate from expected utility maximization at accurate beliefs. At each characteristic $x \in \mathcal{X}$, the smallest $\epsilon(x)$ satisfying Definition 4.1 summarizes how large are the decision maker's violations of expected utility maximization.

Over the class of linear utility functions, the identified set of expected utility costs is characterized by misrankings in the decision maker's choices. This implies tractable characterizations of the total expected utility cost and the share of systematic prediction mistakes in the decision maker's choices.

Theorem 4.1. *Suppose Assumption 3.1 holds and* $0 < \pi_1(x) < 1$ *for all* $x \in \mathcal{X}$. *The decision maker's choices approximately maximize expected utility at some linear utility function and expected utility costs* $\epsilon := \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ *if and only if, for all pairs* $x = (x_I, x_E)$, $x' = (x_I, x'_E)$,

$$\mu_1(x) - \overline{\mu}_0(x') - \epsilon(x) - \epsilon(x') \leq 0. \tag{3}$$

The identified set of expected utility costs equals the set of all $\epsilon = \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ *satisfying* (3).

4.1.1 Bounding the expected utility costs of systematic prediction mistakes

The lower bound on the total expected utility cost of systematic prediction mistakes summarizes how worse off is the decision maker in an expected utility sense relative to hypothetical choices that correctly optimized. By Theorem 4.1, the lower bound is given by the following linear program

$$\underline{\mathcal{E}} := \min_{\epsilon(x)\geq 0:\ x\in\mathcal{X}} \sum_{x\in\mathcal{X}} P(x)\epsilon(x) \tag{4}$$
$$\text{s.t. } \mu_1(x) - \overline{\mu}_0(x') - \epsilon(x) - \epsilon(x') \leq 0 \text{ for all } x = (x_I, x_E), x' = (x_I, x'_E).$$

The lower bound $\underline{\mathcal{E}}$ equals zero if and only if the decision maker's choices are consistent with expected utility maximization at accurate beliefs. For a scalar outcome $Y^* = Y_1^*$, Online Appendix C.2 discusses how the lower bound $\underline{\mathcal{E}}$ can be translated into an equivalent fraction of ex-post errors that arose from the decision maker's systematic prediction mistakes through an accounting exercise. In the pretrial release example, the lower bound $\underline{\mathcal{E}}$ is the judge's total expected utility cost of their systematic prediction mistakes about failure to appear risk, and it can be translated into an equivalent reduction in the fraction of defendants that are released and fail to appear that would produce the same total expected utility cost.

4.1.2 Bounding the share of systematic prediction mistakes

The identified set of expected utility costs further characterizes the share of systematic prediction mistakes in the decision maker's choices. Towards this, I say a subset of characteristics $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ is *rationalizable at accurate beliefs* if there exists a utility function $u \in \mathcal{U}$ and joint distribution $(X, V, C, Y^*) \sim Q(\cdot)$ satisfying Definition 2.2 only over characteristics $x \in \tilde{\mathcal{X}}$. The largest rationalizable subset $\overline{\mathcal{X}}$ of characteristics is then

$$\overline{\mathcal{X}} := \arg\max_{\tilde{\mathcal{X}}\subseteq\mathcal{X}} \sum_{x\in\tilde{\mathcal{X}}} P(x) \text{ s.t. } \tilde{\mathcal{X}} \text{ is rationalizable at accurate beliefs,} \tag{5}$$

and the share of rationalizable decisions is $P(\overline{\mathcal{X}}) := \sum_{x \in \overline{\mathcal{X}}} P(x)$. If the decision maker's choices are consistent with expected utility maximization at accurate beliefs, then $\overline{\mathcal{X}} = \mathcal{X}$ and $P(\overline{\mathcal{X}}) = 1$. The *share of systematic prediction mistakes* in the decision maker's choices is therefore $1 - P(\overline{\mathcal{X}})$. In Online Appendix C.3, I show that the share of systematic prediction mistakes can be characterized by a mixed-integer linear program over the identified set of expected utility costs. This provides a simple summary statistic about the frequency of the decision maker's systematic prediction mistakes.

4.2 Bounding inaccurate beliefs based on characteristics

Misrankings in the decision maker's choices may indicate that their beliefs are *inaccurate* – that is, their implied beliefs do not lie in the identified set $\mathcal{H}(P(\cdot \mid x); \mathcal{B}_x)$. This is a common behavioral hypothesis. Empirical researchers conjecture that judges may systematically mis-predict failure to appear risk based on defendant characteristics (Kleinberg et al., 2018), and the same concern arises in analyses of medical decisions (Currie and Macleod, 2017).

To investigate whether the decision maker's choices maximize expected utility at inaccurate beliefs, I next modify "Data Consistency" in Definition 2.2.

Definition 4.2. The decision maker's choices are *consistent with expected utility maximization at inaccurate beliefs* if there exists some utility function $u \in \mathcal{U}$ and joint distribution $(X, V, C, Y^*) \sim Q(\cdot)$ satisfying (i) Expected Utility Maximization, (ii) Information Set as in Definition 2.2, and

iii. Data Consistency with Inaccurate Beliefs: For all $x \in \mathcal{X}$, there exists $\widetilde{P}_0(\cdot \mid x) \in \mathcal{B}_x$ such that, for all $y^* \in \mathcal{Y}$,

$$Q(c \mid y^*, x)\widetilde{P}(y^* \mid x)Q(x) = \begin{cases} P_1(y^* \mid x)\pi_1(x)P(x) \text{ if } c = 1 \\ \widetilde{P}_0(y^* \mid x)\pi_0(x)P(x) \text{ if } c = 0, \end{cases}$$

where $\widetilde{P}(y^* \mid x) = P_1(y^* \mid x)\pi_1(x) + \widetilde{P}_0(y^* \mid x)\pi_0(x)$.

Definition 4.2 drops the restriction that the decision maker's implied beliefs must lie in the

identified set for the true outcome probabilities. Since it places no direct restrictions on the decision maker's implied prior beliefs $Q(\cdot \mid x)$ nor what gives rise to them, behavior consistent with expected utility maximization at inaccurate beliefs could arise from various behavioral mechanisms. Definition 4.2 therefore nests specific models of belief formation, such as alternative forms of inattention (e.g., Sims, 2003; Gabaix, 2014; Caplin and Dean, 2015) or the use of representativeness heuristics (e.g., Gennaioli and Shleifer, 2010; Bordalo et al., 2016; Bordalo, Gennaioli and Shleifer, 2021). By extending the argument given in the proof of Theorem 3.1, I show that expected utility maximization at inaccurate beliefs and some linear utility function is equivalent to a threshold rule on based on the expectation for $\bar{Y}^*$, reweighed by the likelihood ratio of between the decision maker's implied beliefs and some conditional distribution in the identified set (see Lemma A.2).

In the special case of a binary outcome $Y^* = Y_1^* \in \{0, 1\}$, this implies a bound on the extent to which the decision maker's beliefs overreact or underreact to variation in the characteristics.

Proposition 4.1. *Consider a binary outcome $Y^* = Y_1^* \in \{0, 1\}$, and assume $0 < \pi_1(x) < 1$ for all $x \in \mathcal{X}$. Suppose the decision maker's choices are consistent with expected utility maximization at inaccurate beliefs and some linear utility function at $\widetilde{P}(\cdot \mid x) \in \mathcal{H}(P(\cdot \mid x); \mathcal{B}_x)$ satisfying $0 < \widetilde{P}(1 \mid x) < 1$ for all $x \in \mathcal{X}$. Then, there exists non-negative weights $\omega(y^*; x) \geq 0$ satisfying, for all $x \in \mathcal{X}$,*

$$P_1(1 \mid x) \leq \frac{\omega(0; x) u_{0,1}(x_I)}{\omega(0; x) u_{0,1}(x_I) + \omega(1; x) u_{1,1}(x_I)} \leq \overline{P}_0(1 \mid x), \tag{6}$$

where $\omega(y^*; x) = Q(y^* \mid x,)/\widetilde{P}(y^* \mid x)$ and $Q(y^* \mid x)$, $\widetilde{P}(y^* \mid x)$ are given in Definition 4.2.

Define $\delta(x) := \frac{Q(1|x)/Q(0|x)}{\widetilde{P}(1|x)/\widetilde{P}(0|x)}$ to be the relative odds ratio of the outcome under the decision maker's beliefs relative to the true conditional distribution, and $\tau(x) := \frac{\omega(0;x)u_{0,0}(x_I)}{\omega(0;x)u_{0,0}(x_I)+\omega(1;x)u_{1,1}(x_I)}$ to be the decision maker's reweighed utility threshold. If $\tau(x)$ were known, then $\delta(x)$ could

be backed out as

$$\frac{(1-\tau(x_I,x_E))/\tau(x_I,x_E)}{(1-\tau(x_I,x'_E))/\tau(x_I,x'_E)} = \frac{\delta(x_I,x_E)}{\delta(x_I,x'_E)} \tag{7}$$

for any $x_I \in \mathcal{X}_I$ and $x_E, x'_E \in \mathcal{X}_E$. The ratio $\delta(x_I,x_E)/\delta(x_I,x'_E)$ is an implied prediction mistake, and it summarizes the extent to which the decision maker's beliefs overreact or underreact to variation in the characteristics relative to the true conditional distribution. If the ratio is less than one, then the decision maker's beliefs about the relative probability of $Y_1^* = 1$ versus $Y_1^* = 0$ varies less across the characteristics (x_I, x_E) and (x_I, x'_E) than the true outcome probabilities, and the decision maker's implied beliefs therefore *underreact* across these characteristics. If the ratio is strictly greater than one, then the decision maker's implied beliefs *overreact* across the characteristics in this sense. Since Proposition 4.1 provides an identified set for $\tau(x)$, an identified set for $\delta(x_I,x_E)/\delta(x_I,x'_E)$ can in turn be constructed by computing Equation (7) for each pair $\tau(x_I,x_E), \tau(x_I,x'_E)$ satisfying Equation (6).

5 Do pretrial judges make systematic prediction mistakes?

As an empirical illustration, I apply this econometric framework to analyze the pretrial decisions of judges in New York City. I find that at least 20% of judges in New York City make systematic prediction mistakes in their pretrial release decisions. Under various utility exclusion restrictions, their pretrial decisions are inconsistent with expected utility maximization at accurate beliefs about misconduct outcomes given defendant characteristics.

5.1 Pretrial decisions in New York City

I analyze the pretrial system in New York City, which has been previously studied in Leslie and Pope (2017), Kleinberg et al. (2018) and Arnold, Dobbie and Hull (2022). I observe the universe of all arrests made in New York City between November 1, 2008 and November 1, 2013. This contains information on 1,460,462 cases, of which 758,027 cases were subject to a pretrial release decision. To construct the main estimation sample, I exclude (i)

cases involving non-white and non-black defendants; (ii) cases assigned to judges with fewer than 100 cases; and (iii) cases heard in a court-by-time cell in which there were fewer than 100 cases or only one unique judge, where a court-by-time cell is defined at the assigned courtroom by shift by day of week by month by year level. The main estimation sample consists of 569,256 cases heard by 265 unique judges. I focus on the top 25 judges that heard the most cases over the sample period. These top 25 judges altogether heard 243,118 cases in the main estimation sample, and each judge heard at least 5,000 cases. Online Appendix Table A4 provides descriptive statistics about the main estimation sample and the cases heard by the top 25 judges.

For each case, I observe demographic information about the defendant such as their race, gender, and age, the current charges filed, their criminal record, and their record of pretrial misconduct. I observe a unique identifier for the judge assigned to the case. In each case, the judge decides whether to release the defendant prior to their trial without conditions ("on own recognizance") or set monetary bail conditions. Following Kleinberg et al. (2018) and Arnold, Dobbie and Hull (2022), I code the assigned judge as having released a defendant if either the defendant was released without conditions or the defendant paid cash bail and as having detained a defendant otherwise. I report the robustness of my findings to alternative definitions of the pretrial decision. If the defendant was released, I observe whether the defendant either failed to appear in court or was re-arrested for a new crime. Online Appendix Table A5 reports descriptive statistics broken out by whether the defendant was released or detained.

As discussed in Kleinberg et al. (2018), the New York City pretrial system asks judges to narrowly consider failure to appear risk in deciding whether to release a defendant. I test whether the release decisions of the top 25 judges in New York City maximize expected utility at accurate beliefs about failure to appear risk given defendant characteristics at some linear utility function and private information, assuming that either (i) no defendant characteristics, (ii) the defendant's race, (iii) the defendant's race and age, or (iv) the defendant's race

and charge severity (felony vs. misdemeanor) directly affect the judges' utility function.[5] Nonetheless, judges may consider other outcomes as well, such as whether a defendant would be re-arrested for any new crime or re-arrested for a violent crime. I therefore report the sensitivity of my behavioral conclusions to alternative definitions of the pretrial misconduct outcome. I discretize age into young and older defendants, where older defendants are those older than 25 years.

5.2 Dimension reduction using out-of-sample prediction

A key practical challenge in testing whether judges' release decisions satisfy Proposition 3.2 is that the number of moment inequalities is large, and as a consequence the number of observations per characteristic cell is extremely small. Discretizing all demographic information (e.g., race, age, gender), all current charge information, the prior criminal record, and prior history of pretrial misconduct into binary values produces 134,062 unique characteristic cells with on average 4.24 cases per characteristic cell in the main estimation sample.

To deal with this practical challenge, I test whether there are misrankings in the judges' decisions over a coarsened partition of the characteristics. Define $D\colon \mathcal{X} \to \{1, \ldots, N_d\}$ to partition the characteristics into level sets $\{x\colon D(x) = d\}$. By iterated expectations, if a judge's choices are consistent with expected utility maximization at accurate beliefs and some linear utility function, then there must be no misrankings in their decisions over the partition $D(\cdot)$. Let $\mu_c(x_I, d) := \mathbb{E}[\bar{Y}^* \mid C = c, X_I = x_I, D(X) = d]$ and $\pi_c(x_I, d) := P(C = c \mid X_I = x_I, D(X) = d)$ for $c \in \{0, 1\}$.

Proposition 5.1. *Suppose Assumption 3.1 holds. If the decision maker's choices are consistent with expected utility maximization at accurate beliefs and some linear utility function,*

[5] Empirically the data only contain binary indicators for whether a defendant is black or white. Yet for example research on colorism in the criminal justice system (e.g., King and Johnson, 2016) suggests that judges' prejudices may vary continuously based on defendant skin tone, and more recently Ludwig and Mullainathan (2023) find that defendant facial characteristics influence judges' pretrial decisions. The conjectured exclusion restriction therefore imposes that only the measured race indicator (as well as defendant race or charge severity) may affect judges' utility functions.

then, for all $x_I \in \mathcal{X}_I$

$$\max_{d \in \mathcal{D}_1(x_I)} \mu_1(x_I, d) \leq \min_{x \in \mathcal{D}_0(x_I)} \overline{\mu}_0(x_I, d), \tag{8}$$

where $\mathcal{D}_1(x_I) := \{d \colon \pi_1(x_I, d) > 0\}$ *and* $\mathcal{D}_0(x_I) := \{d \colon \pi_0(x_I, d) > 0\}$.

Provided $N_d \ll |\mathcal{X}_E|$, the number of moment inequalities implied by Equation (8) is drastically reduced and can be tested using methods that rely on an asymptotic normal approximation (Canay and Shaikh, 2017; Molinari, 2020). Searching for misrankings over the coarsened characteristics provides a valid falsification test of whether the decision maker's choices are consistent with expected utility maximization at accurate beliefs.

In a screening decision, a natural choice is to construct the partition $D(\cdot)$ using supervised machine learning methods that predict the outcome $\bar{Y}^*$. Given an estimated prediction function $\hat{f} \colon \mathcal{X} \to [0, K]$, define $D(\cdot)$ by binning the characteristics X into percentiles of predicted risk within each $x_I \in \mathcal{X}_I$. In my empirical analysis, I predict misconduct outcomes among defendants released by all other judges within each $x_I \in \mathcal{X}_I$, defined as either race-by-age cells or race-by-felony charge cells, and partition the characteristics into deciles of predicted risk within each value $x_I \in \mathcal{X}_I$. The prediction function is an ensemble that averages the predictions of an elastic net model and a random forest.

In Online Appendix D, I provide sufficient conditions under which the coarsened inequalities continue to sharply characterize expected utility maximization at accurate beliefs. The characterizations of the expected utility cost of systematic prediction mistakes, the share of systematic prediction mistakes, and bounds on the decision maker's inaccurate beliefs also retain intuitive interpretations after this coarsening step.

5.3 Constructing bounds through the quasi-random assignment of judges

Judges in New York City are quasi-randomly assigned to cases within court-by-time cells defined at the assigned courtroom by shift by day of week by month by year level (see Kleinberg et al., 2018; Arnold, Dobbie and Hull, 2022, for further discussion), which

implies bounds on the conditional failure to appear rate among detained defendants for any particular judge. To verify quasi-random assignment, I conduct balance checks that regress a measure of judge leniency on a rich set of defendant characteristics as well as court-by-time fixed effects that control for the level at which judges are as-if randomly assigned to cases. I measure judge leniency using the leave-one-out release rate among all other defendants assigned to a particular judge (Dobbie, Goldin and Yang, 2018; Arnold, Dobbie and Yang, 2018; Arnold, Dobbie and Hull, 2022). I conduct these balance checks separately within each included characteristic cell (defined by race-by-age cells or race-by-felony-charge cells), reporting the coefficient estimates in Online Appendix Tables A6-A7. In each subsample, the estimated coefficients are economically small in magnitude. A joint F-test fails to reject the null hypothesis of quasi-random assignment for the main estimation sample.

The quasi-random assignment of judges implies bounds on the pretrial misconduct rate among defendants detained by each judge in the top 25. I group judges into quintiles of leniency based on the constructed leniency measure, and define the instrument $Z \in \mathcal{Z}$ to be the leniency quintile of the assigned judge. Applying the results in Online Appendix C.1, the bound on, for example, the failure to appear rate defined as $Y^* = Y_1^* \in \{0, 1\}$ among detained defendants with $X_I = x_I, D(X) = d$ for a particular judge using leniency quintile $\tilde{z} \in \mathcal{Z}$ depends on the quantities $\mathbb{E}[P(C = 1, Y_1^* = 1 \mid X_I = x_I, D(X) = d, Z = \tilde{z}, T) \mid X_I = x_I, D(X) = d]$ and $\mathbb{E}[P(C = 0 \mid X_I = x_I, D(X) = d, Z = \tilde{z}, T) \mid X_I = x_I, D(X) = d]$, where $T \in \mathcal{T}$ denotes the court-by-time cells and the expectation averages over all cases assigned to this particular judge. I tractably model these conditional probabilities as

$$1\{C = 1, Y_1^* = 1\} = \sum_{x_I, d, z} \beta^{c, y^*}_{x_I, d, z} 1\{X_I = x_I, D(X) = d, Z = z\} + \phi_t + \epsilon \tag{9}$$

$$1\{C = 0\} = \sum_{x_I, d, z} \beta^{c}_{x_I, d, z} 1\{X_I = x_I, D(X) = d, Z = z\} + \phi_t + \nu, \tag{10}$$

over all cases in the main estimation sample, where ϕ_t are court-by-time fixed effects. I estimate the relevant quantities by adding the estimated coefficients $\hat{\beta}^{c}_{x_I, d, \tilde{z}}, \hat{\beta}^{c, y^*}_{x_I, d, \tilde{z}}$ to the

average of the respective fixed effects associated with cases heard by the judge within each (x_I, d)-cell.

Figure I plots the observed failure to appear rate among defendants released by the judge that heard the most cases and the resulting estimated upper bound on the failure to appear rate among detained defendants associated with the most lenient quintile of judges at each decile of predicted risk for each race-by-age cell. Testing whether this judge's pretrial release decisions are consistent with expected utility maximization at accurate beliefs about failure to appear risk, then involves checking whether, holding fixed characteristics that directly affect the utility function, all released defendants have a lower probability of failing to appear in court (orange, circles) than the upper bound on the failure to appear rate of all detained defendants (blue, triangles). Online Appendix Figure A1 plots the same quantities for each race-by-felony cell.

5.4 What fraction of judges make systematic prediction mistakes?

By constructing the failure to appear rate among released defendants and the upper bound on the failure to appear rate among detained defendants as in Figure I for each judge in the top 25, I test whether the release decisions of each judge in the top 25 are consistent with expected utility maximization at accurate beliefs about failure to appear risk and some linear utility function satisfying the conjectured exclusion restrictions (Proposition 5.1). I test the moment inequalities that compare the failure to appear rate among released defendants in the top half of the predicted failure to appear risk distribution against the bounds on the failure to appear rate among detained defendants in the bottom half of the predicted failure to appear risk distribution.

The top panel of Table I summarizes the results from testing whether there exists mis-rankings based on failure to appear risk in the release decisions of each judge in the top 25 under various exclusion restrictions. After a multiple hypothesis testing correction that controls the family-wise error rate at the 5% level, the inequalities in Proposition 5.1 are rejected for at least 20% of judges. When both race and age are allowed to directly affect

judges' utility functions, violations imply that the judge's release decisions could not have been generated by any possible discrimination based on the defendant's race and age, any accurate beliefs about failure to appear risk nor variation in private information across defendants. This test allows each judge's utility function to flexibly vary based on defendant race and age, as well as unrestricted heterogeneity in utility functions and private information across judges.

5.4.1 Alternative pretrial misconduct outcomes

If judges' utility functions depend on a richer definition of pretrial misconduct than just failure to appear risk, then the rejections found may reflect mis-specification of the outcome, rather than evidence of systematic prediction mistakes. We may suspect, for example, that judges base their release decisions on whether a defendant would be re-arrested for a new crime and the potential severity of the new crime.

I next define the outcome as $Y^* = (Y_1^*, Y_2^*, Y_3^*)$, where $Y_1^* \in \{0,1\}$ is whether the defendant would fail to appear in court as before, $Y_2^* \in \{0,1\}$ is whether the defendant would be re-arrested for a non-violent crime, and $Y_3^* \in \{0,1\}$ is whether the defendant would be re-arrested for a violent crime (i.e., a violent felony offense, murder, rape or robbery). I now test whether the release decisions of each judge in the top 25 are consistent with expected utility maximization at accurate beliefs about the vector of misconduct outcomes and some linear utility function satisfying the conjectured exclusion restrictions. In this case, the class of linear utility functions places no restrictions on how the relative cost of releasing a defendant that would fail to appear differs from that of a defendant that would be re-arrested for a non-violent or violent crime. By analyzing each judge in the top 25 separately, I further allow any two judges to differentially assess the relative cost of releasing a defendant that would fail to appear versus a defendant that would be re-arrested for a non-violent or violent crime.

The middle panel of Table I summarizes the results from testing whether there exists misrankings based on this vector of misconduct outcomes in the release decisions of each

judge in the top 25 under various exclusion restrictions. I again find that the pretrial release decisions of at least 20% of judges are inconsistent with expected utility maximization at accurate beliefs about failure to appear risk, non-violent crime risk, and violent crime risk and some linear utility function satisfying the conjectured exclusion restrictions.

5.4.2 Incorporating monetary bail conditions

As mentioned earlier, my empirical analysis defined the judge's decision as only a choice between releasing or detaining a defendant. In practice, judges in New York City decide whether to release a defendant without conditions ("on own recognizance"), meaning the defendant is released automatically without bail conditions, or set monetary bail conditions. Consequently, judges could be making two distinct prediction mistakes: first, judges may be systematically mispredicting misconduct outcomes; and second, judges may be systematically mispredicting the ability of defendants to post a specified bail amount.

To investigate this possibility, I instead define the judge's choice to be whether to release the defendant on recognizance ($C = 1$) or set monetary bail conditions ($C = 0$), and I define the outcome Y^* as the vector of whether the defendant would satisfy the monetary bail condition and whether the defendant would fail to appear in court if released. In Online Appendix C.4, I show expected utility maximization at accurate beliefs about bail payment ability and failure to appear risk can again be characterized as a system of moment inequalities. The bottom panel of Table I summarizes the results from testing these resulting moment inequalities under alternative utility exclusion restrictions. I find that the decisions of at least 32% of judges in New York City are inconsistent with expected utility maximization at accurate beliefs about the ability of defendants to post a specified bail amount and failure to appear risk given defendant characteristics.

5.4.3 Alternative bounds on the missing data

Finally, I constructed bounds on the unobserved misconduct rate among defendants detained by each judge in the top 25 using the quasi-random assignment of judges to cases.

Alternative empirical strategies for bounding the missing data are possible, and I next illustrate one such strategy. In particular, suppose that the unobserved failure to appear rate among defendants detained by a particular judge is bounded by the observed failure to appear rate among defendants released by the same judge. That is, for each judge, I apply Proposition 5.1 instead assuming $P(Y_1^* = 1 \mid C = 1, X_I = x_I, D(X) = d) \leq P(Y_1^* = 1 \mid C = 0, X_I = x_I, D(X) = d) \leq (1 + \kappa)P(Y_1^* = 1 \mid C = 0, X_I = x_I, D(X) = d)$ for some chosen parameter $\kappa \geq 0$. Examining how the results change as $\kappa \geq 0$ varies summarizes how alternative assumptions on the magnitude of the missing data problem affects the conclusions about systematic prediction mistakes. In the extreme case with $\kappa = 0$, the unobserved failure to appear rate among defendants detained by a particular judge is point identified.

Figure II reports the fraction of judges in the top 25 for whom we can reject expected utility maximization at accurate beliefs under alternative utility exclusion restrictions and varying the choice of $\kappa \geq 0$. Rationalizing the pretrial release decisions of all judges requires assuming that detained defendants must be at least 5.5 times as risky as released defendants. I therefore continue to find substantial evidence of systematic prediction mistakes in these judges' choices under alternative assumptions about the missing data problem.

5.5 How common and costly are systematic prediction mistakes?

Since the conclusions about the fraction of judges whose decisions are inconsistent with expected utility maximization at accurate beliefs are robust to alternative definitions of pretrial misconduct and bounds on the missing data, I focus on exploring the prediction mistakes about failure to appear risk using bounds on the missing data constructed with the quasi-random assignment of judges to cases. I report analogous results using alternative bounds on the missing data and alternative definitions of pretrial misconduct in Online Appendix E.

While there exists misrankings in the pretrial release decisions of a large fraction of judges, it could be that these systematic prediction mistakes about failure to appear risk only occur over a small subset of defendants or are small in magnitude. To investigate this

possibility, I estimate the bound on the share of systematic prediction mistakes (Section 4.1.2). For each judge whose choices are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level (i.e., an unadjusted rejection), Table II reports the minimum, median and maximum bound on the share of systematic prediction mistakes across these judges. When both defendant race and age may affect judges' utility functions, the median judge makes systematic prediction mistakes on approximately 30% of defendants assigned to them. When both defendant race and whether the defendant was charged with a felony may affect utility, the median judge makes systematic prediction mistakes on approximately 24% of defendants assigned to them.

I next estimate how costly are these uncovered systematic prediction mistakes by calculating the bound on the total expected utility cost of systematic prediction mistakes for each judge whose choices are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level (Section 4.1.1). I translate these estimated expected utility costs into an equivalent reduction in the fraction of defendants that are released and fail to appear in court that would produce the same total expected utility cost to the judge (Online Appendix C.2). For the median judge, this corresponds to an equivalent 9.92 percentage point reduction in the fraction of defendants that are released and fail to appear in court when both defendant race and age are allowed to directly affect utility, and an equivalent 12.1 percentage point reduction when both defendant race and defendant charge severity are allowed to directly affect utility. Taken together, these results indicate that these judges' implied systematic prediction mistakes both occur over a sizeable fraction of defendants and are large in an expected utility sense.

5.6 Bounding prediction mistakes based on defendant characteristics

I next investigate the ways in which the judge's beliefs about failure to appear risk are systematically biased. I apply the identification results in Section 4.2 to bound the extent to

which these judges' implied beliefs overreact or underreact to predictable variation in failure to appear risk based on defendant characteristics. For each judge whose choices are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level, I first construct a 95% joint confidence set for the reweighted utility thresholds $\tau(x_I, d), \tau(x_I, d')$ at the bottom and top deciles of the predicted failure to appear risk distribution using test inversion based on Proposition 4.1. I construct a 95% confidence interval for their implied prediction mistakes $\delta(x_I, d)/\delta(x_I, d')$ between the top decile and bottom deciles of the predicted failure to appear risk distribution by calculating $\frac{(1-\tau(x_I,d))/\tau(x_I,d)}{(1-\tau(x_I,d'))/\tau(x_I,d')}$ for each pair $\tau(x_I, d), \tau(x_I, d')$ in the joint confidence set.

Figure III plots the constructed confidence intervals for the implied prediction mistakes $\delta(x_I, d)/\delta(x_I, d')$ for each judge over the race-and-age cells (see Online Appendix Figure A2 for race-and-felony charge cells) Whenever informative, the confidence intervals highlighted in orange lie everywhere below one, indicating that these judges' are acting as-if their implied beliefs about failure to appear risk underreact to predictable variation in failure to appear risk. These judges are acting as-if they perceive the change in failure to appear risk between defendants in the top decile and bottom decile of predicted risk to be lass then true change in failure to appear risk across these defendants. This could be consistent with judges "regularizing" how their implicit predictions of failure to appear risk respond to variation in the characteristics across these extreme defendants, and may therefore be suggestive of some form inattention (Caplin and Dean, 2015; Gabaix, 2019).

5.7 Which decisions violate expected utility maximization?

As a final qualitative step to investigate why the release decisions of judges in New York City are inconsistent with expected utility maximization at accurate beliefs about failure to appear risk, I report the cells of defendants on which the largest misranking in Proposition 5.1 occurs. This shows which defendants are associated with the largest misrankings in the judges' choices.

Among judges whose choices are inconsistent with expected utility maximization at

accurate beliefs, Table III reports the fraction of judges for whom the maximal studentized misranking occurs over the tails (deciles 1-2, 9-10) or the middle of the predicted failure to appear risk distribution (deciles 3-8) for black and white defendants respectively. All of the largest misrankings in the judges' choices occur over defendants that lie in the tails of the predicted risk distribution. Furthermore, the majority occur over decisions involving black defendants as well. In fact, if the tails of the predicted failure to appear risk distribution are dropped from the original analysis in Section 5.4, all judges' pretrial release decisions over the remaining defendants in the middle of the predicted risk distribution are consistent with expected utility maximization at accurate beliefs. These empirical findings together highlight that release decisions over defendants at the tails of the predicted risk distribution are the primary driver of the documented inconsistencies with expected utility maximization at accurate beliefs about failure to appear risk.

6 The welfare effects of algorithmic decision-making

I finally illustrate the implications of the econometric analysis of systematic prediction mistakes for the design of algorithmic decision systems. I analyze either fully or partially replacing judges in the New York City pretrial system with algorithmic decision rules. The effects of replacing human decisions with algorithms depend on whether judges' make systematic prediction mistakes and if so on which defendants, whether judges are misaligned and optimizing a different objective than the policymaker, and finally whether judges observe any useful private information that is unavailable to the algorithmic decision rules. By allowing for these three competing forces, the analysis of systematic prediction mistakes informs our understanding of the possible tradeoffs between human and algorithmic decision making.

Consider a policymaker with the linear social welfare function $u^*_{1,1}y^*_1c+u^*_{0,1}(1-y^*_1)(1-c)$ for binary outcome $Y^* = Y^*_1 \in \{0,1\}$ and $u^*_{1,1}, u^*_{0,1} \leq 0$. The policymaker evaluates a candidate decision rule $\tilde{\pi}_1(x) \in [0,1]$, which denotes the probability $C = 1$ is chosen given

$X = x$ by the decision rule. Expected social welfare of the decision rule at $x \in \mathcal{X}$ is

$$P(1 \mid x)\tilde{\pi}_1(x)u^*_{1,1} + P(0 \mid x)\tilde{\pi}_0(x)u^*_{0,1}, \tag{11}$$

and total expected social welfare further averages according to the marginal distribution of characteristics. Due to the missing data problem, expected welfare under any candidate decision rule is partially identified. Online Appendix C.5 characterizes the identified set of expected social welfare under alternative candidate decision rules.

I compare expected social welfare under the release decisions of judges in New York City against expected social welfare under particular algorithmic decision rules. Consistent with the stated objectives of the New York City pretrial system, I define the binary outcome to be whether a defendant would fail to appear in court. I specify $u^*_{0,1} = -\tilde{u}/|1 + \tilde{u}|$ and $u^*_{1,1} = -1/|1 + \tilde{u}|$ for some chosen $\tilde{u} \geq 0$, reporting results as $|\tilde{u}|$ varies. For each choice $\tilde{u}$, I construct an algorithmic decision rule that decides whether to release individuals by thresholding a prediction of the probability they would fail to appear at each possible cell of payoff relevant characteristics X_I and each decile of predicted failure to appear risk $D(X)$. The threshold varies based on the parametrization of the social welfare function. By following a threshold rule, there are no misrankings in these algorithmic decisions given its predictions, and in fact it can be shown that these algorithmic decisions are consistent with expected utility maximization at accurate beliefs, albeit with no private information. See Online Appendix C.6 for further details.

I construct 95% confidence intervals for the identified set of expected social welfare under the algorithmic decision rule and the judge's observed released decisions, reporting the ratio of worst-case expected social welfare under the algorithmic decision rule against the judge's observed release decisions. I conduct this exercise for each judge over the race-by-age cells, reporting the median, minimum and maximum gain across judges. Online Appendix E.3 reports results over the race-by-felony charge cells.

6.1 Replacing judges who make systematic prediction mistakes

I compare the release decisions of judges who make systematic prediction mistakes about failure to appear risk against algorithmic decision rules that fully replace them over all defendants. These judges made systematic prediction mistakes over defendants in the tails of the predicted failure to appear risk distribution. Over the remaining defendants, however, their choices are consistent with expected utility maximization at accurate beliefs about failure to appear risk, some private information, and a utility function that varied based on defendant race and age.

The policymaker's comparison between fully replacing judges with algorithmic decision rules versus the judges' release decisions is driven by three forces. First, the algorithmic decision rules may improve decisions by correcting these judges' systematic prediction mistakes over the tails of the predicted failure to appear risk distribution. Second, the algorithmic decision rules may improve decisions by correcting possible misalignment between the policymaker's social welfare function $u^*_{0,1}, u^*_{1,1}$ which does not directly depend on defendant characteristics, and these judges' utility functions $u_{1,1}(x_I), u_{0,1}(x_I)$ over the remaining defendants. Finally, in the other direction, these judges may observe useful private information over the remaining defendants that is unavailable to the algorithm.

In order to trace out these competing effects, Figure IVa compares the worst-case expected social welfare under the algorithmic decision rules against the release decisions of each judge, varying the policymaker's social welfare function $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|, u^*_{1,1} = -1/|1+\tilde{u}|$. Worst-case expected social welfare under the algorithmic decision rules is strictly larger than worst-case expected social welfare under these judges' decisions for most values of the policymaker's social welfare function. Intriguingly for $|\tilde{u}| \in [0.3, 0.8]$, the algorithmic decision rules either lead to no improvement or strictly lower worst-case expected social welfare relative to these judges' decisions. Online Appendix Figure A3 plots the comparison of worst-case expected social welfare by the race of the defendant, highlighting that these costs of the algorithmic decision rule are particularly large over white defendants. For these values, the

cost of misalignment with the policymaker's objective may be outweighed by the value of these judges' private information over the remaining defendants, and so it is costly to fully replace them with algorithmic decision rules. To further investigate this hypothesis, Online Appendix Figure A4 compares the release rates of the algorithmic decision rules that directly optimize the social planner's objective function against the observed release rates of these judges. Indeed, the judges' observed release rates are most similar to the algorithmic decision rules' release rates precisely over the values of the social welfare function where these judges' decisions dominate the algorithmic decision rules.

The preceding econometric analysis, however, suggests that it would be most valuable to replace these judges with algorithmic decision rules precisely over the defendants at the tails of the predicted failure to appear risk distribution. Over these defendants, there exists *no* private information nor utility function that could rationalize these judges' choices at any accurate beliefs about failure to appear risk. I therefore next compare these judges' observed release decisions against algorithmic decision rules that only replace them over defendants in the tails of the predicted failure to appear risk distribution but otherwise defers to the judges on all remaining defendants. Such a triage rule may reap the benefits of correcting systematic prediction mistakes where they occur while avoiding the potential loss of valuable private information on the remaining defendants. Figure IVb reports the welfare effects of partially replacing judges with algorithmic decision rules as the policymaker's social welfare function varies. Strikingly, I find that the algorithmic decision rule that corrects systematic prediction mistakes weakly dominates the observed release decisions of judges, no matter the value of the social welfare function. For some parametrizations, the algorithmic decision rule leads to 20% improvements in worst-case social welfare relative to the observed release decisions of these judges.

These comparisons of judges' decisions against algorithmic decision rules that either fully or partially replace judges highlight how the analysis of systematic prediction mistakes informs the design of algorithmic decision systems. Recent machine learning methods attempt

to directly optimize whether particular decisions should be made by directly by an algorithm or instead should be deferred to an existing decision maker (e.g., Madras, Pitassi and Zemel, 2018; Raghu et al., 2019; Wilder, Horvitz and Kamar, 2020). By simultaneously allowing for systematic prediction mistakes, possible misalignment and private information, the econometric analysis directly informs us about these specific behavioral mechanisms which govern the effects of replacing decision makers with algorithmic decision rules.

6.2 Replacing judges who do not make systematic prediction mistakes

I next compare the release decisions of judges whose choices were found to be consistent with expected utility maximization at accurate beliefs about failure to appear risk against algorithmic decision rules that fully replace them over all defendants. The policymaker's comparison between these judges' release decisions and fully replacing them with algorithmic decision rules now only depends on two forces: first, the algorithmic decision rules may correct possible misalignment between the policymaker's objective and these judges' utility functions; and second, these judges may observe useful private information that is unavailable to the algorithms.

Online Appendix Figure A5 reports the welfare effects of replacing these judges with algorithmic decision rules, varying $|\tilde{u}|$. Replacing these judges' release decisions may strictly lower worst-case expected social welfare for a range of social welfare functions. For these values, the cost of potential misalignment may be outweighed by the value of these judges' private information, leading to better decisions than the algorithmic decision rules on average from the policymaker's perspective. Online Appendix Figure A6 compares the algorithm's release rates against the observed release rates of these judges, showing that their observed release rates are most similar to the algorithmic decision rule over the values of the social welfare function where the status quo dominates the counterfactual. Altogether, the effects of fully replacing a decision maker whose decisions are consistent with expected utility max-

imization at accurate beliefs again depends on the tradeoff between the value of their private information against the degree to which they are misaligned with the policymaker. Exploring how policymakers may empirically design optimal delegation rules to such decision makers (e.g., see Frankel, 2021) in light of this econometric framework is an important question for future work.

7 Conclusion

This paper develops an econometric framework for testing whether a decision maker makes systematic prediction mistakes in high stakes settings like pretrial release and many others. I characterized expected utility maximization behavior, where the decision maker maximizes some utility function at accurate beliefs as well as some private information. I developed a statistical test for whether the decision maker makes systematic prediction mistakes and methods for estimating the ways in which their predictions are systematically biased. Analyzing the New York City pretrial system, I found that a substantial fraction of judges make systematic prediction mistakes about pretrial misconduct risk given defendant characteristics.

Machine learning based models are now increasingly used in high-stakes decisions, and specific behavioral hypotheses about decision makers often underlie their design. The effects of fully or partially replacing decision makers with algorithmic decision rules depend on whether decision makers make systematic prediction mistakes and if so on which decisions, whether decision makers are misaligned and optimizing a different objective than the policymaker, and finally whether decision makers observe useful private information that is unavailable to the algorithmic decision rules. By combining quasi-experimental variation with formal identification analysis, researchers can explore the empirical content of expected utility maximization at accurate beliefs, and thereby inform our understanding of the possible tradeoffs between human and algorithmic decision making.

Of course, machine learning based models are also frequently used to *inform* decision

makers by providing them with a recommended decision or risk prediction (e.g., see Stevenson and Doleac, 2022; Agarwal et al., 2023; Albright, 2023; Angelova, Dobbie and Yang, 2023; Grimon and Mills, 2023, among many others). The design of such algorithmic decision aids requires a behavioral understanding of the human decision maker "in-the-loop." Indeed, recent work begins to shed empirical light on how exactly human decision makers form beliefs and respond to algorithms in these settings. For example, by examining variation in estimated testing probabilities, Mullainathan and Obermeyer (2022) provide suggestive evidence that doctors may simultaneously fail to pay attention to some relevant patient characteristics yet overweight other salient characteristics. Judges may pay attention to facial characteristics of defendants (Ludwig and Mullainathan, 2023), and mistakenly overrule algorithmic recommendations as they overweight recent salient cases (Angelova, Dobbie and Yang, 2023), and respond differentially to algorithmic recommendations based on defendant race (Albright, 2023). While I explored how to robustly test for and measure deviations from the rational benchmark of expected utility maximization at accurate beliefs, the next step is to explore the formal implications of alternative behavioral hypotheses such as inattention (e.g., Sims, 2003; Gabaix, 2014; Caplin and Dean, 2015) as well as forms of salience or representativeness (e.g., Gennaioli and Shleifer, 2010; Bordalo et al., 2016; Bordalo, Gennaioli and Shleifer, 2021). Such results are the missing econometric link needed to operationalize behavioral insights into the design of algorithmic decision systems. Exploiting these high-stakes settings as rich laboratories for behavioral economics is an important, policy-relevant agenda at the intersection of economic theory and microeconometrics, motivated by the application of machine learning in these empirical settings.

Ashesh Rambachan

Massachusetts Institute of Technology

Department of Economics

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): 3730–3764.

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.

Albright, Alex. 2023. "The Hidden Effects of Algorithmic Recommendations."

Andrews, Isaiah, Jonathan Roth, and Ariel Pakes. 2023. "Inference for Linear Conditional Moment Inequalities." 6.

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 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.

Autor, David H., and David Scarborough. 2008. "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments." *The Quarterly Journal of Economics*, 123(1): 219–277.

Beaulieu-Jones, Brett, Samuel G. Finlayson, Corey Chivers, Irene Chen, Matthew McDermott, Jaz Kandola, Adrian V. Dalca, Andrew Beam, Madalina Fiterau, and Tristan Naumann. 2019. "Trends and Focus of Machine Learning Applications for Health Research." *JAMA Network Open*, 2(10): e1914051–e1914051.

Becker, Gary. 1957. *The Economics of Discrimination.* University of Chicago Press.

Belloni, Alexandre, Federico Bugni, and Victor Chernozhukov. 2018. "Subvector Inference in Partially Identified Models with Many Moment Inequalities." arXiv preprint, arXiv:1806.11466.

Bergemann, Dirk, and Stephen Morris. 2016. "Bayes correlated equilibrium and the comparison of information structures in games." *Theoretical Economics*, 11: 487–522.

Bergemann, Dirk, and Stephen Morris. 2019. "Information Design: A Unified Perspective." *Journal of Economic Literature*, 57(1): 44–95.

Bergemann, Dirk, Benjamin Brooks, and Stephen Morris. 2022. "Counterfactuals with Latent Information." *American Economic Review*, 112(1): 343–368.

Blattner, Laura, and Scott T. Nelson. 2021. "How Costly is Noise?" arXiv preprint, arXiv:arXiv:2105.07554.

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

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2021. "Salience." NBER Working Paper Series No. 29274.

Camerer, Colin. F., and Eric J. Johnson. 1997. "The Process-Performance Paradox in Expert Judgement." In *Research on Judgment and Decision Making: Currents, Connections, and Controversies.* , ed. W. M. Goldstein and R. M. Hogarth. New York: Cambridge University Press.

Canay, Ivan A., and Azeem M. Shaikh. 2017. "Practical and Theoretical Advances in Inference for Partially Identified Models." *Advances in Economics and Econometrics: Eleventh World Congress*, , ed. Bo Honoré, Ariel Pakes, Monika Piazzesi and Larry Samuelson Vol. 2, 271–306. Cambridge University Press.

Canay, Ivan, Magne Mogstad, and Jack Mountjoy. 2020. "On the Use of Outcome Tests for Detecting Bias in Decision Making." NBER Working Paper No. 27802.

Caplin, Andrew. 2021. "Economic Data Engineering." NBER Working Paper Series No. 29378.

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

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

Caplin, Andrew, Dàniel Csaba, John Leahy, and Oded Nov. 2020. "Rational Inattention, Competitive Supply, and Psychometrics." *The Quarterly Journal of Economics*, 135(3): 1681–1724.

Caplin, Andrew, Daniel J Martin, and Philip Marx. 2022. "Modeling Machine Learning." National Bureau of Economic Research Working Paper 30600.

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.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 134–148.

Cox, Gregory, and Xiaoxia Shi. 2022. "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models." *The Review of Economic Studies*.

Currie, Janet, and W. Bentley Macleod. 2017. "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians." *Journal of Labor Economics*, 35(1): 1–43.

Dawes, Robyn M. 1971. "A case study of graduate admissions: Application of three principles of human decision making." *American Psychologist*, 26(2): 180–188.

Dawes, Robyn M. 1979. "The robust beauty of improper linear models in decision making." *American Psychologist*, 34(7): 571–582.

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. "Clinical Versus Actuarial Judgment." *Science*, 249(4899): 1668–1674.

de Chaisemartin, Clement. 2017. "Tolerating Defiance? Local Average Treatment Effects Without Monotonicity." *Quantitative Economics*, 8(2): 367–396.

Dobbie, Will, and Crystal S. Yang. 2021. "The US Pretrial System: Balancing Individual Rights and Public Interests." *Journal of Economic Perspectives*, 35(4): 49–70.

Dobbie, Will, and Crystal Yang. 2019. "Proposals for Improving the U.S. Pretrial System." The Hamilton Project.

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

Einav, Liran, Mark Jenkins, and Jonathan Levin. 2013. "The impact of credit scoring on consumer lending." *Rand Journal of Economics*, 44(2): 249—-274.

Erel, Isil, Lea H. Stern, Chenhao Tan, and Michael S. Weisbach. 2019. "Selecting Directors Using Machine Learning." NBER Working Paper Series No. 24435.

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

Frankel, Alexander. 2021. "Selecting Applicants." *Econometrica*, 89(2): 615–645.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." *The Journal of Finance*, 77(1): 5–47.

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *The Quarterly Journal of Economics*, 129(4): 1661–1710.

Gabaix, Xavier. 2019. "Behavioral Inattention." In *Handbook of Behavioral Economics: Applications and Foundations*. Vol. 2, , ed. B. Douglas Bernheim, Stefano DellaVigna and David Laibson, 261–343. North Holland.

Gennaioli, Nicola, and Andrei Shleifer. 2010. "What Comes to Mind." *The Quarterly Journal of Economics*, 125(4): 1399–1433.

Grimon, Marie-Pascale, and Christopher Mills. 2023. "The Impact of Algorithmic Tools on Child Protection: Evidence from a Randomized Controlled Trial."

Grove, W. M., D. H. Zald, B. S. Lebow, B. E. Snitz, and C. Nelson. 2000. "Clinical versus mechanical prediction: A meta-analysis." *Psychological Assessment*, 12(1): 19–30.

Gualdani, Christina, and Shruti Sinha. 2020. "Identification and Inference in Discrete Choice Models with Imperfect Information." arXiv preprint, arXiv:1911.04529.

Handel, Benjamin, and Joshua Schwartzstein. 2018. "Frictions or Mental Gaps: What's Behind the Information We (Don't) Use and When Do We Care?" *Journal of Economic Perspectives*, 32(1): 155–178.

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

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2020. "Simple rules to guide expert classifications." *Journal of the Royal Statistical Society Series A*, 183(3): 771–800.

King, Ryan D., and Brian D. Johnson. 2016. "A Punishing Look: Skin Tone and Afrocentric Features in the Halls of Justice." *American Journal of Sociology*, 122(1): 90–124.

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

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review: Papers and Proceedings*, 105(5): 491–495.

Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review*, 96(3): 863–876.

Kuncel, Nathan R., David M. Klieger, Brian S. Connelly, and Deniz S Ones. 2013. "Mechanical Versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis." *Journal of Applied Psychology*, 98(6): 1060—1072.

Leslie, Emily, and Nolan G. Pope. 2017. "The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from New York City Arraignments." *The Journal of Law and Economics*, 60(3): 529–557.

Li, Danielle, Lindsey Raymond, and Peter Bergman. 2020. "Hiring as Exploration." NBER Working Paper Series No. 27736.

Ludwig, Jens, and Sendhil Mullainathan. 2023. "Machine Learning as a Tool for Hypothesis Generation." National Bureau of Economic Research Working Paper 31017.

Madras, David, Toniann Pitassi, and Richard Zemel. 2018. "Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer." arXiv preprint, arXiv:1711.06664.

Magnolfi, Lorenzo, and Camilla Roncoroni. 2021. "Estimation of Discrete Games with Weak Assumptions on Information."

Manski, Charles F. 1994. "The Selection Problem." In *Advances in Econometrics: Sixth World Congress*. Vol. 1, , ed. Christopher Sims, 143–170. Cambridge University Press.

Martin, Daniel, and Phillip Marx. 2022. "A Robust Test of Prejudice for Discrimination Experiments." *Management Science*, 68(6): 3975–4753, iv–v.

Molinari, Francesca. 2020. "Microeconometrics with Partial Identification." In *Handbook of Econometrics*. Vol. 7, 355–486.

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

Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine." *The New England Journal of Medicine*, 375(13): 1216–9.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." 469–481.

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

Rambachan, Ashesh, and Jens Ludwig. 2021. "Empirical Analysis of Prediction Mistakes in New York City Pretrial Data." University of Chicago Crime Lab Technical Report.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2021. "An Economic Approach to Regulating Algorithms." NBER Working Paper Series No. 27111.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika*, 63(3): 581–592.

Sims, Christopher A. 2003. "Implications of rational inattention." *Journal of Monetary Economics*, 50(3): 665–690.

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

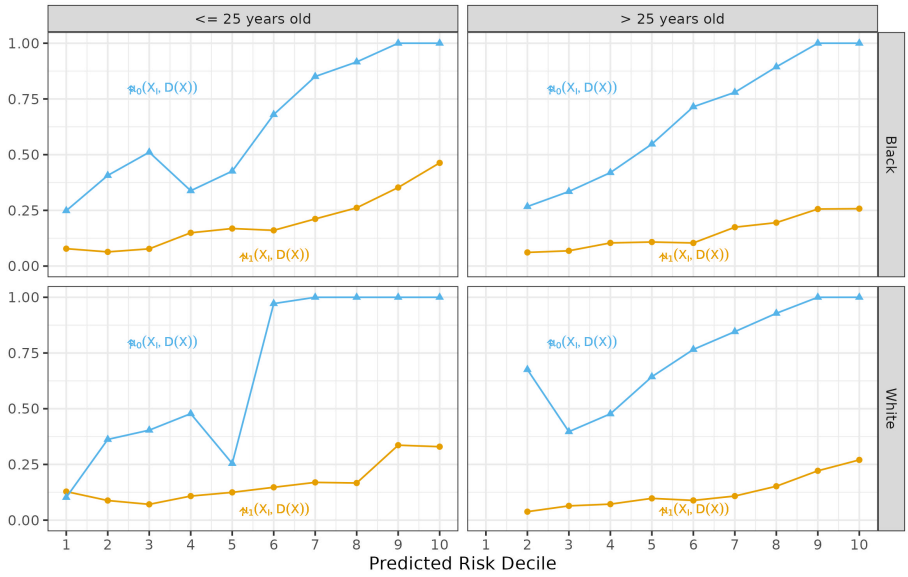
Stevenson, Megan, and Jennifer Doleac. 2022. "Algorithmic Risk Assessment in the Hands of Humans."

Syrgkanis, Vasilis, Elie Tamer, and Juba Ziani. 2018. "Inference on Auctions with Weak Assumptions on Information." arXiv preprint, arXiv:1710.03830.

Wilder, Bryan, Eric Horvitz, and Ece Kamar. 2020. "Learning to Complement Humans." 1526–1533. International Joint Conferences on Artificial Intelligence Organization.

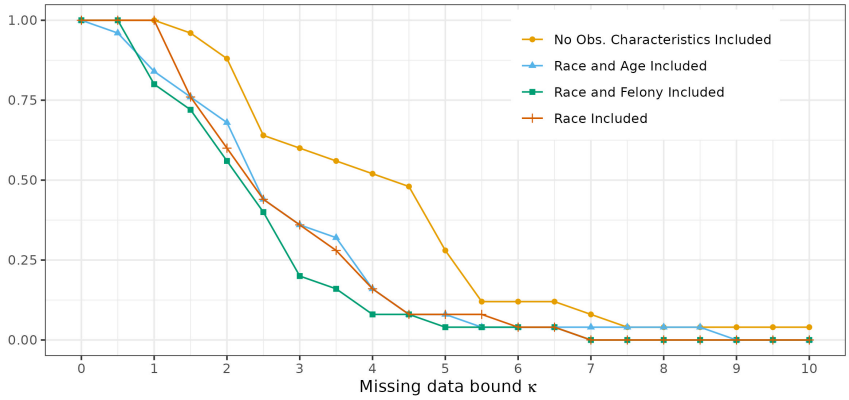
Figures

Figure I: Judge-specific failure to appear rate among released defendants and bound on the failure to appear rate among detained defendants.



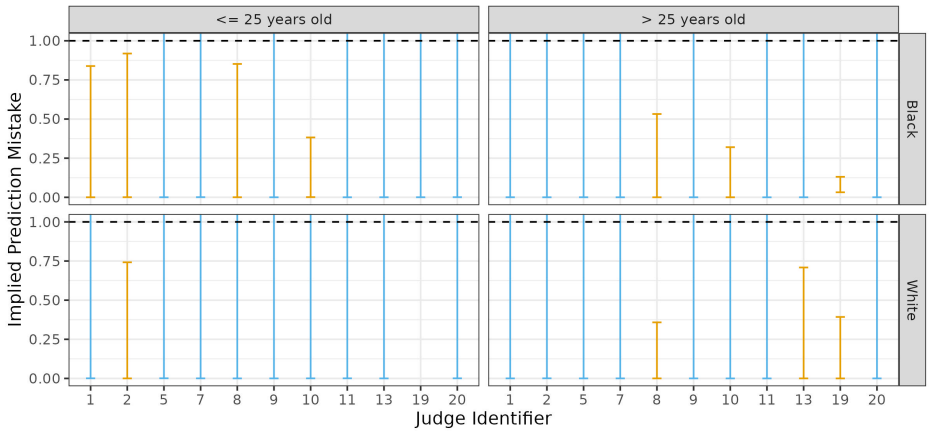
Notes: This figure plots the failure to appear rate among released defendants (orange, circles) and the bounds on the failure to appear rate among detained defendants based on the judge leniency instrument (blue, triangles) at each decile of predicted failure to appear risk and race-by-age cell for the judge that heard the most cases in the main estimation sample. See Section 5.3 for further estimation details on these bounds.

Figure II: Fraction of judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs.



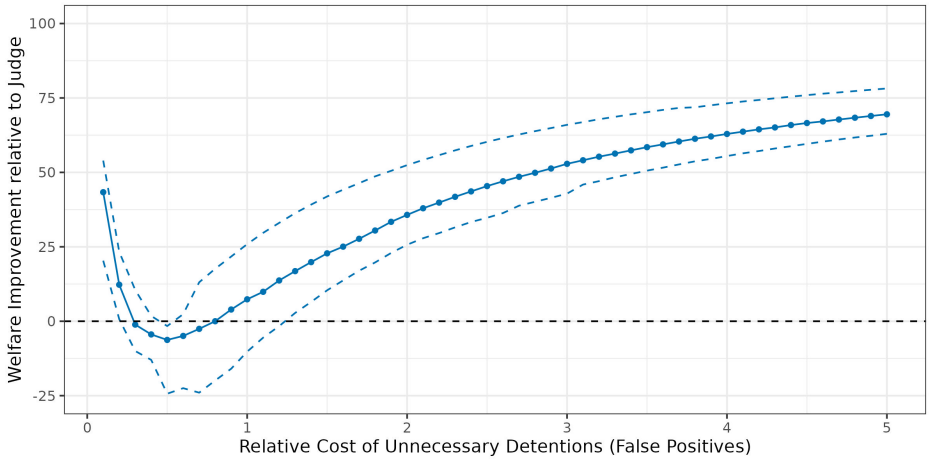
Notes: This figure summarizes the results for testing whether the release decisions of each judge in the top 25 are consistent with expected utility maximization behavior at a linear utility function $u(c, y^*; x_I)$ that (i) does not depend on any observable characteristics, (ii) depends on the defendant's race, (iii) depends on both the defendant's race and age, and (iv) depends on both the defendant's race and whether the defendant was charged with a felony offense. Bounds on the failure to appear rate among detained defendants are constructed using bounds using the alternative bounding strategy discussed in Section 5.4.3 for $\kappa = \{0, 1, \ldots, 10\}$. I test the moment inequalities using the conditional least-favorable hybrid test developed in Andrews, Roth and Pakes (2023). I estimate the variance-covariance matrix of the failure to appear rate among released defendants using the bootstrap conditional on the included characteristics X_I and predicted risk decile $D(X)$. The adjusted rejection rate reports the fraction of rejections after a multiple hypothesis testing correction that controls the family-wise error rate at the 5% level.

Figure III: Judge-specific bounds on prediction mistakes between predicted failure to appear risk deciles.

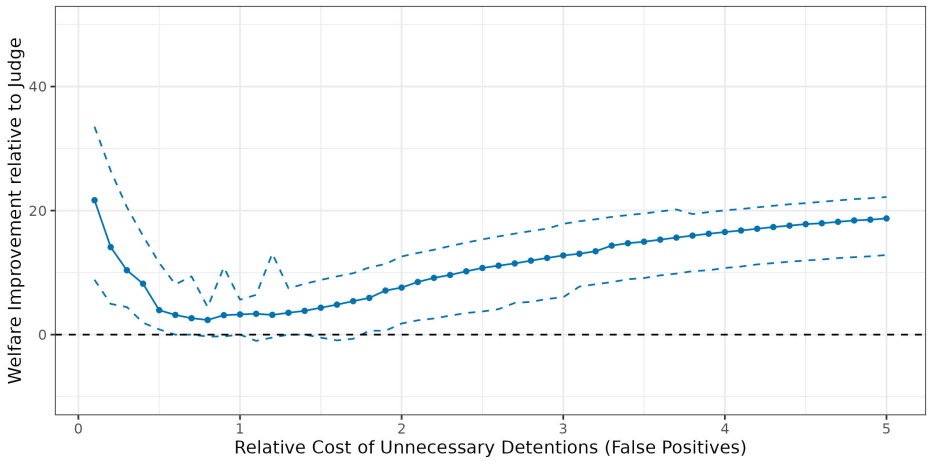


Notes: This figure plots the 95% confidence interval on the implied prediction mistake $\delta(x_I, d)/\delta(x_I, d')$ between the top decile d and bottom decile d' of the predicted failure to appear risk distribution for each judge in the top 25 whose pretrial release decisions are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level (the "unadjusted rejection rate" in the top panel of Table II) and each race-by-age cell. When informative, the confidence intervals highlighted in orange show that judges under-react to predictable variation in failure to appear risk from the highest to the lowest decile of predicted failure to appear risk (i.e., the estimated bounds lie below one). See Section 4.2 for theoretical details on the implied prediction mistake and Section 5.6 for the estimation details.

Figure IV: Comparison of algorithmic decision rule against judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs.



(a) Algorithmic decision rule that fully replaces judges.



(b) Algorithmic decision rule that corrects prediction mistakes.

Notes: This figure reports the change in worst-case expected social welfare under two algorithmic decisions rules against the release decisions of judges whose pretrial release decisions are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level (the "unadjusted rejection rate" in the top panel of Table II). Panel (a) reports the comparison for an algorithmic decision rule that fully replaces judges over all decisions. Panel (b) reports the comparison for an algorithmic decision rule that only replaces judges over the tails of the predicted risk distribution. The x-axis plots the relative social welfare cost of detaining a defendant that would not fail to appear in court $|\tilde{u}|$ (i.e., an unnecessary detention), where $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$, $u^*_{1,1} = -1/|1+\tilde{u}|$. The solid line plots the median change and the dashed lines report the minimum and maximum change across judges that make systematic prediction mistakes. See Section 6 for further details.

Tables

Table I: Estimated fraction of judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs.

	Utility Functions $u(c, y^*; x_I)$			
	No Characteristics	Race	Race + Age	Race + Felony Charge
A) Failure to appear				
Unadjusted Rejection Rate	48%	48%	48%	56%
Adjusted rejection rate	24%	24%	20%	32%
B) Misconduct outcomes				
Adjusted rejection rate	24%	24%	20%	32%
C) Release on recognizance				
Adjusted rejection rate	32%	32%	32%	52%

Notes: This table summarizes the results for testing whether each judge's pretrial release decisions are consistent with expected utility maximization at accurate beliefs and some linear utility function $u(c, y^*; x_I)$ that (i) do not depend on any defendant characteristics, (ii) depend on the defendant's race, (iii) depend on both the defendant's race and age, and (iv) depend on both the defendant's race and whether the defendant was charged with a felony offense. Panel A tests for misrankings based on failure to appear risk alone, (see Section 5.4). Panel B alternatively defines the outcome $Y^* = (Y_1^*, Y_2^*, Y_3^*) \in \{0,1\}^3$ as whether the defendant would fail to appear in court Y_1^*, be re-arrested for a non-violent crime Y_2^*, and be re-arrested for a violent crime Y_3^* upon release (see Section 5.4.1). Panel C tests whether the "release on recognizance" vs monetary bail decisions of judges are consistent with expected utility maximization behavior at accurate beliefs (see Section 5.4.3). In each specification, bounds on the missing outcomes among detained defendants are constructed using the judge leniency instrument (see Section 5.3). I test the moment inequalities using the conditional least-favorable hybrid test developed in Andrews, Roth and Pakes (2023). I estimate the variance-covariance matrix of moments using the bootstrap conditional on the included characteristics X_I, predicted risk decile $D(X)$ and leniency quintile instrument Z. The "adjusted rejection rate" reports the fraction of rejections across all judges in the top 25 after a multiple hypothesis testing correction that controls the family-wise error rate at the 5% level.

Table II: Share of systematic prediction mistakes among judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs.

	Utility Functions $u(c, y^*; x_I)$	
	Race and Age	Race and Felony Charge
Unadjusted Rejection Rate	48%	56%
Prediction Mistake Share		
Minimum	6.30%	11.88%
Median	30.87%	24.45%
Maximum	42.26%	45.78%

Notes: This table summarizes the estimated bound on the share of systematic prediction mistakes among judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs and linear utility functions that depend on both the defendant's race and age as well as the defendant's race and whether the defendant was charged with a felony. Among judges' whose pretrial release decisions are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level (the "unadjusted rejection rate" in the top panel of Table I), I compute the optimal value of the sample analogue to the optimization program (21). See Section 4.1.2 for theoretical details on the bound for the share of systematic prediction mistakes and Section 5.5 for the estimation details.

Table III: Largest misrankings among judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs.

	Utility Functions $u(c, y^*; x_I)$	
	Race and Age	Race and Felony Charge
Unadjusted Rejection Rate	48%	56%
White Defendants		
Middle Deciles	0%	0%
Tail Deciles	25%	7.14%
Black Defendants		
Middle Deciles	0%	0%
Tail Deciles	75%	92.85%

Notes: This table summarizes the location of the largest (studentized) misranking in Proposition 5.1 among judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs and linear utility functions that depend on both the defendant's race and age as well as the defendant's race and whether the defendant was charged with a felony. Among judges whose pretrial release decisions are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level (the "unadjusted rejection rate" in the top panel of Table I), I report the fraction of judges for whom the largest studentized misranking occurs among white and black defendants on tail deciles (deciles 1-2, 9-10) and middle deciles (3-8) of predicted failure to appear risk.

Online Appendix for "Identifying Prediction Mistakes in Observational Data"

Ashesh Rambachan

May 1, 2024

Contents

A Omitted proofs

A.1 Section 3: identifying systematic prediction mistakes in screening decisions

A.1.1 Proof of Theorem 3.1

Lemma A.1. *The decision maker's choices are consistent with expected utility maximization at some linear utility function if and only if there exists some linear utility function satisfying*

i. $\mu_1(x_I, x_E) \leq \sum_{k=1}^K |u_{0,k}(x_I)|$ *for all* $(x_I, x_E) \in \mathcal{X}_I \times \mathcal{X}_E$ *with* $\pi_1(x_I, x_E) > 0$,

ii. $\sum_{k=1}^K |u_{0,k}(x_I)| \leq \overline{\mu}_0(x_I, x_E)$ *for all* $(x_I, x_E) \in \mathcal{X}_I \times \mathcal{X}_E$ *with* $\pi_0(x_I, x_E) > 0$.

Proof. I apply Theorem B.1 to a screening decision with a binary choice. Theorem B.1 requires $\mu_1(x_I, x_E) \leq \sum_{k=1}^K |u_{0,k}(x_I)|$ for all $(x_I, x_E) \in \mathcal{X}_I \times \mathcal{X}_E$ with $\pi_1(x_I, x_E) > 0$, and $\sum_{k=1}^K |u_{0,k}(x_I)| \leq \mu_0(x_I, x_E)$ for all $(x_I, x_E) \in \mathcal{X}_I \times \mathcal{X}_E$ with $\pi_0(x_I, x_E) > 0$. Applying the sharp bound $\mu_0(x) \leq \max_{\widetilde{P}(\cdot|x) \in \mathcal{B}_x} \mu_0(x) := \overline{\mu}_0(x)$ then delivers the result. □

By Lemma A.1, the decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a linear utility function $u(c, y^*; x_I)$ satisfying, for all $x_I \in \mathcal{X}_I$,

$$\max_{x_E \in \Pi_1(x_I)} \mu_1(x_I, x_E) \leq \sum_{k=1}^K |u_{0,k}(x_I)| \leq \min_{x_E \in \Pi_0(x_I)} \overline{\mu}_0(x_I, x_E)$$

The inequalities in Theorem 3.1 are immediate. The identified set of linear utility functions is then given by all linear utility functions (Definition 3.1) that satisfy the above display. □

A.1.2 Proof of Proposition 3.1

Recall $\bar{Y}^* := \sum_{k=1}^K Y_k^*$, $\mu(x, z) := \mathbb{E}[\bar{Y}^* \mid X = x, Z = z]$ and $\mu_c(x, z) := \mathbb{E}[\bar{Y}^* \mid C = c, X = x, Z = z]$ for $c \in \{0, 1\}$ Under Assumption 3.2, $\mu(x, z) = \mu(x, \tilde{z}) = \mu(x)$ for all $x \in \mathcal{X}$ and $z, \tilde{z} \in \mathcal{Z}$. Using that $Y_k^* \in [0, 1]$ for all $k = 1, \ldots, K$ and applying worst-case bounds (e.g., Manski, 1994), $\mu(x, z)$ is sharply bounded by

$$\mu_1(x, \tilde{z})\pi_1(x, \tilde{z}) \leq \mu(x, z) \leq K\pi_0(x, \tilde{z}) + \mu_1(x, \tilde{z})\pi_1(x, \tilde{z}).$$

The result follows by (i) writing $\mu(x, z) = \mu_0(x, z)\pi_0(x, z) + \mu_1(x, z)\pi_1(x, z)$ via iterated expectations, (ii) taking the maximum, minimum of the lower, upper bounds respectively over $\tilde{z} \in \mathcal{Z}$, and (iii) re-arranging. □

A.2 Section 4: characterizing systematic prediction mistakes in screening decisions

A.2.1 Proof of Theorem 4.1

I apply Theorem B.2 to a screening decision with a binary choice over the class of linear utility functions. For all $(x_I, x_E) \in \mathcal{X}_I \times \mathcal{X}_E$ with $\pi_1(x_I, x_E) > 0$, Theorem B.2 requires

$\mu_1(x_I, x_E) - \epsilon(x_I, x_E) \leq \sum_{k=1}^K |u_{0,k}(x_I)|$. For all $(x_I, x_E) \in \mathcal{X}_I \times \mathcal{X}_E$ with $\pi_0(x_I, x_E) > 0$, Theorem B.2 requires $\sum_{k=1}^K |u_{0,k}(x_I)| \leq \overline{\mu}_0(x_I, x_E) + \epsilon(x_I, x_E)$. Putting these together, it follows that the decision maker's choices approximately maximize expected utility at $\epsilon = \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ if and only if, for all $x_I \in \mathcal{X}_I$,

$$\max_{x_E \in \mathcal{X}_E} \{\mu_1(x_I, x_E) - \epsilon(x_I, x_E)\} \leq \min_{x'_E \in \mathcal{X}_E} \{\overline{\mu}_0(x_I, x'_E) + \epsilon(x_I, x'_E)\}.$$

This is equivalent to, for all $x_I \in \mathcal{X}_I$,

$$\mu_1(x_I, x_E) - \overline{\mu}_0(x_I, x'_E) - \epsilon(x_I, x_E) - \epsilon(x_I, x'_E) \leq 0 \text{ for all } x_E, x'_E \in \mathcal{X}_E.$$

□

A.2.2 Proof of Proposition 4.1

To prove this result, we first establish the following Lemma.

Lemma A.2. *Suppose Assumption 3.1 holds, $\widetilde{P}(\cdot \mid x) > 0$ for all $\widetilde{P}(\cdot \mid x) \in \mathcal{H}(P(\cdot \mid x); \mathcal{B}_x)$, $x \in \mathcal{X}$, and $0 < \pi_1(x) < 1$ for all $x \in \mathcal{X}$. The decision maker's choices are consistent with expected utility maximization at inaccurate beliefs and some linear utility function if and only if there exists $\widetilde{P}_0(\cdot \mid x) \in \mathcal{B}_x$ and non-negative weights $\omega(y^*; x)$ satisfying, for all $x_I \in \mathcal{X}_I$,*

$$\max_{\tilde{x}_E \in \Pi_1(x_I)} \mathbb{E}_{\widetilde{P}}[\omega_1(Y^*; X)\bar{Y}^* \mid C = 1, X = (x_I, \tilde{x}_E)] \leq \min_{\tilde{x}_E \in \Pi_0(x_I)} \mathbb{E}_{\widetilde{P}}[\omega_0(Y^*; X)\bar{Y}^* \mid C = 0, X = (x_I, \tilde{x}_E)]$$

and, for all $x \in \mathcal{X}$, $\mathbb{E}_{\widetilde{P}}[\omega(Y^*; X) \mid X = x] = 1$, where $\omega_1(y^*; x) = \omega(y^*; x)/\mathbb{E}_{\widetilde{P}}[\omega(Y^*; X) \mid C = 1, X = x]$, $\omega_0(y^*; x) = \omega(y^*; x)/\mathbb{E}_{\widetilde{P}}[\omega(Y^*; X) \mid C = 0, X = x]$ and $\mathbb{E}_{\widetilde{P}}[\cdot]$ is the expectation under the joint distribution $(X, C, Y^*) \sim \widetilde{P}(\cdot)$ defined in the proof.

Proof. This is an immediate consequence of applying Theorem B.3 to a binary choice, screening decision over the class of linear utility functions. For all $x \in \mathcal{X}$ with $\pi_1(x) > 0$, Theorem B.3 requires

$$\mathbb{E}_{\widetilde{P}}[\omega(Y^*; X)\bar{Y}^* \mid C = 1, X = x] \leq \mathbb{E}_{\widetilde{P}}[\omega(Y^*; X) \mid C = 1, X = x] \left(\sum_{k=1}^K |u_{0,k}(x_I)| \right).$$

Defining $\omega_1(Y^*; X) = \omega(Y^*; X)/\mathbb{E}_{\widetilde{P}}[\omega(Y^*; X) \mid C = 1, X = x]$, this can be equivalently written as

$$\mathbb{E}_{\widetilde{P}}[\omega_1(Y^*; X)\bar{Y}^* \mid C = 1, X = x] \leq \sum_{k=1}^K |u_{0,k}(x_I)|.$$

Similarly, for all $x \in \mathcal{X}$ with $\pi_0(x) > 0$, Theorem B.3 requires

$$\sum_{k=1}^K |u_{0,k}(x_I)| \leq \mathbb{E}_{\widetilde{P}}[\omega_0(Y^*; X)\bar{Y}^* \mid C = 0, X = x],$$

where $\omega_0(Y^*; X) = \omega(Y^*; X)/\mathbb{E}_{\widetilde{P}}[\omega(Y^*; X) \mid C = 0, X = x]$. It therefore follows that the

decision maker's choices are consistent with expected utility maximization at inaccurate beliefs if and only if there exists a linear utility function, $\widetilde{P}(\cdot \mid x) \in \mathcal{B}_x$ for all $x \in \mathcal{X}$ and non-negative weights $\omega(y^*; x)$ satisfying, for all $x_I \in \mathcal{X}_I$,

$$\max_{\tilde{x}_E \in \Pi_1(x_I)} \mathbb{E}_{\widetilde{P}}[\omega_1(Y^*; X)\bar{Y}^* \mid C = 1, X = x] \leq \sum_{k=1}^{K} |u_{0,k}(x_I)|,$$
$$\sum_{k=1}^{K} |u_{0,k}(x_I)| \leq \min_{\tilde{x}_E \in \Pi_0(x_I)} \mathbb{E}_{\widetilde{P}}[\omega_0(Y^*; X)\bar{Y}^* \mid C = 0, X = x]$$

and, for all $x \in \mathcal{X}$, $\mathbb{E}_{\widetilde{P}}[\omega(Y^*; X) \mid X = x] = 1$. □ □

Under the stated conditions, the necessity statement in Theorem B.3 implies that, for all $x \in \mathcal{X}$,

$$\omega(1; x)u_{1,1}(x_I)P_1(1 \mid x) \geq \omega(0; x)u_{0,1}(x_I)P_1(0 \mid x),$$
$$\omega(0; x)u_{0,1}(x_I)\widetilde{P}_0(0 \mid x) \geq \omega(1; x)u_{1,1}(x_I)\widetilde{P}_0(1 \mid x),$$

where $\omega(y^*; x) = \frac{\widetilde{Q}(y^*|x)}{\widetilde{P}(y^*|x)}$. Re-arranging these inequalities, we observe that

$$P_1(1 \mid x) \leq \frac{\omega(0; x)u_{0,1}(x_I)}{\omega(0; x)u_{0,1}(x_I) + \omega(1; x)u_{1,1}(x_I)} \leq \widetilde{P}_0(1 \mid x).$$

The result follows by applying the bounds on $\widetilde{P}_0(1 \mid x)$. □

A.3 Section 5: do pretrial release judges make prediction mistakes?

A.3.1 Proof of Proposition 5.1

This is an immediate consequence of Lemma B.1 to a screening decision and iterated expectations. Since the decision maker's choices are consistent with expected utility maximization, by Lemma B.1, there exists some linear utility function u and $\widetilde{P}_0(\cdot \mid x) \in \mathcal{B}_x$ such that her choices satisfy, for all $x \in \mathcal{X}$,

$$\sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X = x)u(1, y^*; x_I) \geq \sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X = x)u(0, y^*; x_I)$$
$$\sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X = x)u(0, y^*; x_I) \geq \sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X = x)u(1, y^*; x_I),$$

where $\widetilde{P}(Y^* = y^*, C = 0 \mid X = x) = \widetilde{P}_0(y^* \mid x)\pi_0(x)$. Therefore, her choices satisfy, for all $d \in \{1, \ldots, N_d\}$,

$$\sum_{x_E:\, D(x_I, x_E) = d} \sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X = (x_I, x_E)) \frac{P(X_E = x_E \mid X_I = x_I)}{P(D(X_I, X_E) = d \mid X_I = x_I)} u(1, y^*; x_I) \geq$$

$$\sum_{x_E:\, D(x_I,x_E)=d} \sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X = (x_I, x_E)) \frac{P(X_E = x_E \mid X_I = x_I)}{P(D(X_I, X_E) = d \mid X_I = x_I)} u(0, y^*; x_I)$$

and

$$\sum_{x_E:\, D(x_I,x_E)=d} \sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X = (x_I, x_E)) \frac{P(X_E = x_E \mid X_I = x_I)}{P(D(X_I, X_E) = d \mid X_I = x_I)} u(0, y^*; x_I) \geq$$

$$\sum_{x_E:\, D(x_I,x_E)=d} \sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X = (x_I, x_E)) \frac{P(X_E = x_E \mid X_I = x_I)}{P(D(X_I, X_E) = d \mid X_I = x_I)} u(1, y^*; x_I).$$

These can equivalently be written as

$$\sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X_I = x_I, D(X) = d) u(1, y^*; x_I) \geq$$

$$\sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X_I = x_I, D(X) = d) u(0, y^*; x_I)$$

and

$$\sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X_I = x_I, D(X) = d) u(0, y^*; x_I) \geq$$

$$\sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X_I = x_I, D(X) = d) u(1, y^*; x_I).$$

The result then follows by the same argument as the proof of Theorem 3.1. □

B Expected utility maximization in treatment assignment problems

In the main text, I made two simplifying assumptions for exposition: (i) the decision maker only faced two choices; and (ii) the decision maker's choice did not have a direct causal effect on the outcome. I now relax both assumptions, and analyze *treatment assignment problems* in which the decision maker selects one of many treatments for each individual. This nests the main text analysis of screening decisions as a special case.

B.1 Setting and behavioral model

The decision maker selects a choice $c \in \{c_1, \ldots, c_J\}$ for each individual. Each individual is summarized by characteristics $x \in \mathcal{X}$ and a vector of potential outcomes. The *potential outcome* $y_j := y(c_j) \in \mathcal{Y}$ is the outcome that would occur if the decision maker selects choice c_j. Let $\vec{y} = (y_1, \ldots, y_J) \in \mathcal{Y}^J$ denote the vector of potential outcomes associated with each choice, and $\vec{y}_{-j}$ is the vector of all potential outcomes except for the potential outcome associated with choice c_j.

The random vector $(X, C, \vec{Y}) \sim P(\cdot)$ summarizes the joint distribution of the characteristics, the decision maker's choices and potential outcomes across all individuals. I assume the characteristics and outcome have finite support, and there exists $\delta > 0$ such that

$P(x) := P(X = x) \geq \delta$ for all $x \in \mathcal{X}$. This nests the main text as a special case if we further assume (i) choice is binary $c \in \{0, 1\}$; and (ii) choices do not have a causal effect on the outcome with $y_1 = y^*$, $y_0 = 0$.

We observe the potential outcome associated with the decision maker's choice, where $Y := \sum_{j=1}^{J} Y_j 1\{C = c_j\}$. We observe the conditional potential outcome probabilities $P(Y_j = y \mid C = c_j, X = x)$ for $j = 1, \ldots, J$, but not the counterfactual potential outcome probabilities $P(Y_k = y \mid C = c_j, X = x)$ for $j \neq k$. As notation, let $P_j(\vec{y} \mid x) := P(\vec{Y} = \vec{y} \mid C = c_j, X = x)$, and $P_j(\cdot \mid x) \in \Delta(\mathcal{Y}^J)$ denote the conditional distribution $\vec{Y} \mid C = c_j, X = x$. Let $\pi_j(x) := P(C = c_j \mid X = x)$ denote the generalized propensity score for each $c_j \in \{c_1, \ldots, c_J\}$.

For each choice c_j and characteristic $x \in \mathcal{X}$, I assume there exists a known subset $\mathcal{B}_{j,x} \subseteq \Delta(\mathcal{Y}^J)$ such that $P_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ and $\sum_{\vec{y}_{-j}} \widetilde{P}_j((\vec{y}_{-j}, y_j) \mid x) = P(Y_j = y_j \mid C = c_j, X = x)$ for all $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ and $y_j \in \mathcal{Y}$. Denote the identified set for $P(\cdot \mid x) := P(\vec{Y} \mid X = x)$ as $\mathcal{H}(P(\cdot \mid x); \mathcal{B}_x)$, where $\mathcal{B}_x := \{\mathcal{B}_{j,x} \colon j = 1, \ldots, J\}$.

Definition B.1. The *utility function* $u \colon \{c_1, \ldots, c_J\} \times \mathcal{Y}^J \times \mathcal{X}_I$ specifies the payoff associated with each choice, vector of potential outcomes, and characteristics $x_I \in \mathcal{X}_I$. Let $\mathcal{U}$ denote the feasible set of utility functions specified by the researcher.

Definition B.2. The decision maker's choices are *consistent with expected utility maximization* in a treatment assignment problem if there exists $u \in \mathcal{U}$ and $(X, V, C, \vec{Y}) \sim Q(\cdot)$ satisfying

i. Expected Utility Maximization: For all $c_j \in \{c_1, \ldots, c_J\}$, $(x, v) \in \mathcal{X} \times \mathcal{V}$ such that $Q(c_j \mid x, v) > 0$,

$$\mathbb{E}_Q\left[u(c_j, \vec{Y}; X_I) \mid X = x, V = v\right] \geq \mathbb{E}_Q\left[u(c', \vec{Y}; X_I) \mid X = x, V = v\right]$$

for all $c' \neq c_j$, where $\mathbb{E}_Q[\cdot]$ denotes the expectation under $Q(\cdot)$.

ii. Information Set: $C \perp\!\!\!\perp \vec{Y} \mid \{X, V\}$ under $Q(\cdot)$.

iii. Data Consistency: For all $c_j \in \{c_1, \ldots, c_J\}$, $x \in \mathcal{X}$, there exists $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ satisfying, for all $\vec{y} \in \mathcal{Y}^J$,

$$Q(x, c_j, \vec{y}) = \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)P(x).$$

The *identified set of utility functions* is the set of $u \in \mathcal{U}$ such that there exists $(X, V, C, \vec{Y}) \sim Q(\cdot)$ satisfying (i)-(iii).

B.2 Characterization results

The decision maker's choices in a treatment assignment problem are consistent with expected utility maximization behavior if and only if there exists a utility function that satisfies a set of stochastic revealed preference inequalities.

Theorem B.1. *The decision maker's choices in a treatment assignment problem are consistent with expected utility maximization behavior if and only if there exists $u \in \mathcal{U}$ and $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ for all $c_j \in \{c_1, \ldots, c_J\}$, $x \in \mathcal{X}$ such that*

$$\mathbb{E}_Q\left[u(c_j, \vec{Y}; X_I) \mid C = c_j, X = x\right] \geq \mathbb{E}_Q\left[u(c', \vec{Y}; X_I) \mid C = c_j, X = x\right] \tag{12}$$

for all $c' \neq c_j$ whenever $\pi_j(x) > 0$, where $(X, C, \vec{Y}) \sim Q(\cdot)$ is given by $Q(x, c_j, \vec{y}) = \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)P(x)$. The identified set of utility functions is the set of all utility functions $u \in \mathcal{U}$ such that there exists $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ for all $c_j \in \{c_1, \ldots, c_J\}$, $x \in \mathcal{X}$ satisfying (12).

Theorem B.1 provides a necessary and sufficient characterization of expected utility maximization that only involves the data and the bounds on the conditional potential outcome probabilities.

As in Section 3 of the main text, I next analyze the testable implications of expected utility maximization behavior for a binary choice $c \in \{0, 1\}$ over linear utility functions.

Assumption B.1 (Vector-valued outcome)**.** For some $K \geq 1$, each potential outcome satisfies $y_j = y(c_j) = (y_1(c_j), \ldots, y_K(c_j)) \in [0, 1]^K$ for $j = 1, \ldots, J$.

Under Assumption B.1, the class of linear utility functions is the set of utility function satisfying $u(c, \vec{y}; x_I) = \sum_{k=1}^K y_k - u_{0,k}(x_I)c$, where $u_{0,k}(x_I) \geq 0$ for all $x_I \in \mathcal{X}_I$. As in the main text, define $\Pi_c(x_I) := \{x_E \colon \pi_1(x_E, x_I) > 0\}$ for $c \in \{0, 1\}$. Let $\bar{Y}(c) := \sum_{k=1}^K Y_k(c)$, $\mu_c(x) := \mathbb{E}[\bar{Y}(1) - \bar{Y}(0) \mid C = c, X = x]$ for each $c \in \{0, 1\}$, and $\underline{\mu}_0(x) := \min_{\widetilde{P}_0(\cdot \mid x) \in \mathcal{B}_{0,x}} \mu_0(x)$, $\overline{\mu}_1(x) := \max_{\widetilde{P}_1(\cdot \mid x) \in \mathcal{B}_{1,x}} \mu_1(x)$.

Corollary B.1. *The decision maker's choices are consistent with expected utility maximization at some linear utility function if and only if, for all $x_I \in \mathcal{X}_I$,*

$$\max_{x_E \in \Pi_0(x_I)} \underline{\mu}_0(x_I, x_E) \leq \min_{x_E \in \Pi_1(x_I)} \overline{\mu}_1(x_I, x_E). \tag{13}$$

The identified set of linear utility functions equals the set of all utility functions satisfying, for all $x_I \in \mathcal{X}_I$, $u(c, \vec{y}; x_I) = \sum_{k=1}^K y_k - u_{0,k}(x_I)c$ with $u_{0,k}(x_I) \geq 0$ and

$$\max_{x_E \in \Pi_0(x_I)} \underline{\mu}_0(x_I, x_E) \leq \sum_{k=1}^K |u_{0,k}(x_I)| \leq \min_{x_E \in \Pi_1(x_I)} \overline{\mu}_1(x_I, x_E). \tag{14}$$

Corollary B.1 immediately implies two negative results about the identification of systematic prediction mistakes in treatment assignment problems that parallel Corollary 3.1 in the main text for screening decisions.

Corollary B.2. *The decision maker's choices are consistent with expected utility maximization behavior at some linear utility function $u(c, \vec{y}; x_I) = \sum_{k=1}^K y_k + u_{0,k}(x_I)c$ if either:*

(i) All characteristics affect utility (i.e., $\mathcal{X} = \mathcal{X}_I$) and $\underline{\mu}_0(x_I) \leq \overline{\mu}_1(x_I)$ for all $x_I \in \mathcal{X}_I$.

(ii) The researcher's bounds on the conditional potential outcome probabilities are uninformative, meaning, for both $c \in \{0, 1\}$ and all $x \in \mathcal{X}$, $\mathcal{B}_{c,x}$ equals the set of all $\widetilde{P}_c(\cdot \mid x)$ satisfying $\sum_{y_{\tilde{c}} \in \mathcal{Y}} \widetilde{P}_c(y_c, y_{\tilde{c}} \mid x) = P_c(y_c \mid x)$ for all $y_c \in \mathcal{Y}$.

B.3 Approximate expected utility maximization in treatment assignment problems

I characterize conditions under which the decision maker's choices approximately maximize expected utility at accurate beliefs in a treatment assignment problem. The definition of approximate expected utility maximization behavior in the main text (Definition 4.1) generalizes naturally.

Definition B.3. The decision maker's choices are consistent with *approximate expected utility maximization* in a treatment assignment problem if there exists $u \in \mathcal{U}$, expected utility costs $\epsilon(x) \geq 0$ for all $x \in \mathcal{X}$, and $(X, V, C, \vec{Y}) \sim Q(\cdot)$ satisfying:

i. Approximate Expected Utility Maximization: For all $c \in \{0, 1\}$, $c' \neq c$, $(x, v) \in \mathcal{X} \times \mathcal{V}$ such that $Q(c \mid x, v) > 0$,

$$\mathbb{E}_Q \left[u(c, \vec{Y}; X_I) \mid X = x, V = v \right] \geq \mathbb{E}_Q \left[u(c', \vec{Y}; X_I) \mid X = x, V = v \right] - \epsilon(x),$$

and (ii) Information Set, (iii) Data Consistency as defined in Definition B.2. The *identified set of expected utility costs* is the set of $\epsilon = \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ such that there exists $u \in \mathcal{U}$ and $(X, V, C, \vec{Y}) \sim Q(\cdot)$ satisfying (i)-(iii).

Theorem B.2. *The decision maker's choices are consistent with approximate expected utility maximization if and only if there exists $u \in \mathcal{U}$, $\epsilon(x) \geq 0$ for all $x \in \mathcal{X}$, and $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ for all $c_j \in \{c_1, \ldots, c_J\}$, $x \in \mathcal{X}$ satisfying*

$$\mathbb{E}_Q \left[u(c_j, \vec{Y}; X_I) \mid C = c, X = x \right] \geq \mathbb{E}_Q \left[u(c', \vec{Y}; X_I) \mid C = c, X = x \right] - \epsilon(x) \tag{15}$$

for all $c' \neq c_j$ whenever $\pi_j(x) > 0$, where $(X, C, \vec{Y}) \sim Q(\cdot)$ is given by $Q(x, c, \vec{y}) = \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)P(x)$.

Corollary B.3. *Consider treatment assignment problem with binary choice, and suppose $0 < \pi_1(x) < 1$ for all $x \in \mathcal{X}$. The decision maker's choices approximately maximize expected utility at some linear utility function and expected utility costs $\epsilon(x) \geq 0$ for all $x \in \mathcal{X}$ if and only if, for all pairs $x = (x_I, x_E)$, $x' = (x_I, x'_E)$,*

$$\underline{\mu}_0(x) - \overline{\mu}_1(x') - \epsilon(x) - \epsilon(x') \leq 0. \tag{16}$$

The identified set of expected utility costs equals the set of all $\epsilon = \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ satisfying (16).

B.4 Expected utility maximization at inaccurate beliefs in treatment assignment problems

I finally characterize conditions under which the decision maker's choices are consistent with expected utility maximization at inaccurate beliefs in a treatment assignment problem.

Definition B.4. The decision maker's choices are *consistent with expected utility maximization at inaccurate beliefs* in a treatment assignment if there exists $u \in \mathcal{U}$ and $(X, V, C, \vec{Y}) \sim$

$Q(\cdot)$ satisfying (i) Expected Utility Maximization, (ii) Information Set as in Definition B.2, and

iii. Data Consistency with Inaccurate Beliefs: For all $x \in \mathcal{X}$, there exists $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ for each $j = 1, \ldots, J$ such that, for all $\vec{y} \in \mathcal{Y}^J$ and $c_j \in \{c_1, \ldots, c_J\}$,

$$Q(c_j \mid \vec{y}, x)\widetilde{P}(\vec{y} \mid x)Q(x) = \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)P(x),$$

where $\widetilde{P}(\vec{y} \mid x) = \sum_{j=1}^{J} \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)$.

Theorem B.3. *Assume $\widetilde{P}(\cdot \mid x) > 0$ for all $\widetilde{P}(\cdot \mid x) \in \mathcal{H}(P(\cdot \mid x); \mathcal{B}_x)$ and $x \in \mathcal{X}$. The decision maker's choices are consistent with expected utility maximization at inaccurate beliefs in a treatment assignment problem if and only if there exists $u \in \mathcal{U}$, $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ for all $j = 1, \ldots, J$ and $x \in \mathcal{X}$, and non-negative weights $\omega(\vec{y}; x)$ satisfying*

i. For all $c_j \in \{c_1, \ldots, c_J\}$, $x \in \mathcal{X}$ with $\pi_j(x) > 0$, $c' \neq c_j$

$$\mathbb{E}_{\widetilde{P}}\left[\omega(\vec{Y}; X)u(c_j, \vec{Y}; X_I) \mid C = c_j, X = x\right] \geq \mathbb{E}_{\widetilde{P}}\left[\omega(\vec{Y}; X)u(c', \vec{Y}; X_I) \mid C = c_j, X = x\right]$$

ii. For all $x \in \mathcal{X}$, $\mathbb{E}_{\widetilde{P}}[\omega(\vec{Y}; X) \mid X = x] = 1$

where $\mathbb{E}_{\widetilde{P}}[\cdot]$ is the expectation under $(X, C, \vec{Y}) \sim \widetilde{P}(\cdot)$ defined as $\widetilde{P}(x, c_j, \vec{y}) = \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)P(x)$.

B.5 Proofs of characterization results for treatment assignment problems

B.5.1 Proof of Theorem B.1

Proof. I prove the following Lemma, and then show it implies Theorem B.1.

Lemma B.1. *The decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a utility function $u \in \mathcal{U}$, $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ for each $c_j \in \{c_1, \ldots, c_J\}$ and $x \in \mathcal{X}$ satisfying*

$$\sum_{\vec{y} \in \mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)u(c', \vec{y}; x_I)$$

for all $x \in \mathcal{X}$, $c \in \{c_1, \ldots, c_J\}$, $c' \neq c_j$,

Proof of Lemma B.1: Necessity Suppose that the decision maker's choices are consistent with expected utility maximization at some utility function U and joint distribution $(X, V, C, \vec{Y}) \sim Q$.

First, I show that if the decision maker's choices are consistent with expected utility maximization behavior at some utility function u, joint distribution $(X, V, C, \vec{Y}) \sim Q$ in which private information has support $\mathcal{V}$, then her choices are also consistent with expected utility maximization behavior at some finite support private information. I show this for the

case where $J = 2$, and the argument generalizes to $J > 2$ at the expense of more cumbersome notation.

Partition the original signal space $\mathcal{V}$ into the subsets $\mathcal{V}_{\{1\}}, \mathcal{V}_{\{2\}}, \mathcal{V}_{\{1,2\}}$, which collect together the signals $v \in \mathcal{V}$ at which the decision maker strictly prefers $C = c_1$, strictly prefers $C = c_2$ and is indifferent between $C = c_1, C = c_2$ respectively. Define the coarsened signal space $\widetilde{\mathcal{V}} = \{v_{\{1\}}, v_{\{2\}}, v_{\{1,2\}}\}$ and coarsened private information $\widetilde{V} \in \widetilde{\mathcal{V}}$ as

$$\begin{aligned}
\widetilde{Q}(\widetilde{V} = v_{\{1\}} \mid \vec{Y} = \vec{y}, X = x) &= Q(V \in \mathcal{V}_{\{1\}} \mid \vec{Y} = \vec{y}, X = x) \\
\widetilde{Q}(\widetilde{V} = v_{\{2\}} \mid \vec{Y} = \vec{y}, X = x) &= Q(V \in \mathcal{V}_{\{2\}} \mid \vec{Y} = \vec{y}, X = x) \\
\widetilde{Q}(\widetilde{V} = v_{\{1,2\}} \mid \vec{Y} = \vec{y}, X = x) &= Q(V \in \mathcal{V}_{\{1,2\}} \mid \vec{Y} = \vec{y}, X = x).
\end{aligned}$$

Define $\widetilde{Q}(C = c_1 \mid \widetilde{V} = v_{\{1\}}, X = x) = 1$, $\widetilde{Q}(C = c_2 \mid \widetilde{V} = v_{\{2\}}, X = x) = 1$ and

$$\widetilde{Q}(C = c_2 \mid \widetilde{V} = v_{\{1,2\}}, X = x) = \frac{Q(C = c_2, V \in \mathcal{V}_{\{1,2\}} \mid X = x)}{Q(V \in \mathcal{V}_{\{1,2\}} \mid W = w, X = x)}.$$

Define the coarsened expected utility representation by the utility function u and the random vector $(X, \widetilde{V}, C, \vec{Y}) \sim \widetilde{Q}$, where $\widetilde{Q}(x, v, c, \vec{y}) = Q(x, \vec{y})\widetilde{Q}(\widetilde{v} \mid x, \vec{y})\widetilde{Q}(c \mid x, \widetilde{v})$. The information set and expected utility maximization conditions are satisfied by construction. Data consistency is satisfied since it is satisfied at the original private information $V \in \mathcal{V}$. To see this, notice that for all $(x, \vec{y}) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Y}^2$

$$\begin{aligned}
&P(C = c_2, \vec{Y} = \vec{y} \mid X = x) = \\
&Q(C = c_2, V = \mathcal{V}, \vec{Y} = \vec{y} \mid X = x) = \\
&Q(C = c_2, V \in \mathcal{V}_{\{2\}}, \vec{Y} = \vec{y} \mid X = x) + Q(C = 1, V \in \mathcal{V}_{\{1,2\}}, \vec{Y} = \vec{y} \mid X = x) = \\
&\widetilde{Q}(C = c_2, \widetilde{V} = v_2, \vec{Y} = \vec{y} \mid X = x) + \widetilde{Q}(C = 1, \widetilde{V} = v_{1,2}, \vec{Y} = \vec{y} \mid X = x) = \\
&\sum_{\widetilde{v} \in \widetilde{\mathcal{V}}} \widetilde{Q}(C = c_2, \widetilde{V} = \widetilde{v}, \vec{Y} = \vec{y} \mid X = x) = \widetilde{Q}(C = c_2, \vec{Y} = \vec{y} \mid X = x).
\end{aligned}$$

The same argument applies to $P(C = c_1, \vec{Y} = \vec{y} \mid X = x)$. For the remainder of the necessity proof, it is therefore without loss to assume private information $V \in \mathcal{V}$ has finite support.

I next show that if there exists an expected utility representation for the decision maker's choices, then the stated inequalities in Lemma B.1 are satisfied by adapting the necessity argument given the "no-improving action switches inequalities" in Caplin and Martin (2015). Suppose that the decision maker's choices are consistent with expected utility maximization at utility function $u \in \mathcal{U}$ and joint distribution $(X, V, C, \vec{Y}) \sim Q$. Then, for each $c_j \in \{c_1, \ldots, c_J\}$ and $(x, v) \in \mathcal{X} \times \mathcal{V}$,

$$Q(c_j \mid x, v) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c_j, \vec{y}; x_I) \right) \geq Q(c_j \mid x, v) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c', \vec{y}; x_I) \right)$$

holds for all $c_j \neq c'$. If $Q(c_j \mid x, v) = 0$, this holds trivially. If $Q(c_j \mid x, v) > 0$, this holds

through the expected utility maximization condition. Multiply both sides by $Q(v \mid x)$ to arrive at

$$Q(c_j \mid x, v)Q(v \mid x)\left(\sum_{\vec{y}\in\mathcal{Y}^J} Q(\vec{y} \mid x, v)u(c_j, \vec{y}; x_I)\right) \geq Q(c_j \mid x, v)Q(v \mid x)\left(\sum_{\vec{y}\in\mathcal{Y}^J} Q(\vec{y} \mid x, v)u(c', \vec{y}; x_I)\right).$$

Next, use information set to write $Q(c_j, \vec{y} \mid x, v) = Q(\vec{y} \mid x, v)Q(c_j \mid x, v)$ and arrive at

$$Q(v \mid x)\left(\sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x, v)u(c_j, \vec{y}; x_I)\right) \geq Q(v \mid x)\left(\sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x, v)u(c', \vec{y}; x_I)\right).$$

Finally, we use $Q(c_j, \vec{y}, v \mid x) = Q(c_j, \vec{y} \mid x, v)Q(v \mid x)$ and then further sum over $v \in \mathcal{V}$ to arrive at

$$\sum_{\vec{y}\in\mathcal{Y}^J}\left(\sum_{v\in\mathcal{V}} Q(v, c_j, \vec{y} \mid x)\right)u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y}\in\mathcal{Y}^J}\left(\sum_{v\in\mathcal{V}} Q(v, c_j, \vec{y} \mid x)\right)u(c', \vec{y}; x_I)$$
$$\sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x)u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x)u(c', \vec{y}; x_I).$$

The inequalities in Lemma B.1 follow from an application of data consistency.

Proof of Lemma B.1: Sufficiency As notation, let $\mathcal{C} := \{c_1, \ldots, c_J\}$. To establish sufficiency, I show that if the conditions in Lemma B.1 hold, then private information $v \in \mathcal{V}$ can be constructed that recommends choices $c \in \mathcal{C}$ and an expected utility maximizer would find it optimal to follow these recommendations as in the sufficiency argument in Caplin and Martin (2015) for the "no-improving action switches" inequalities.

Towards this, suppose that the conditions in Lemma B.1 are satisfied at some $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ for all $j = 1, \ldots, J$, $x \in \mathcal{X}$. As notation, let $v \in \mathcal{V} := \{1, \ldots, 2^J\}$ index all possible subsets in the power set $2^{\mathcal{C}}$.

For each $x \in \mathcal{X}$, define $\mathcal{C}_x := \{c_j \colon \pi_j(x) > 0\} \subseteq \mathcal{C}$ to be the set of choices selected with positive probability, and partition $\mathcal{C}_x$ into subsets that have identical conditional outcome probabilities. There are $\bar{V}_x \leq |\mathcal{C}_x|$ such subsets. Each subset of this partition of $\mathcal{C}_x$ is a subset in the power set $2^{\mathcal{C}}$, and so I may associate each subset in this partition with its index $v \in \mathcal{V}$. Denote the conditional outcome probability associated with the subset labelled v by $P(\cdot \mid v, x) \in \Delta(\mathcal{Y}^J)$. Finally, define $Q(\vec{y} \mid x) = \sum_{j=1}^{J} \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)$.

Define $V \in \mathcal{V}$ according to

$$Q_V(v \mid x) = \sum_{c_j \colon P_j(\cdot|x)=P(\cdot|v,x)} \pi_j(x) \text{ if } v \in \mathcal{V}_x,$$
$$Q_V(v \mid \vec{y}, x) = \begin{cases} \frac{P(\vec{y}|v,x)Q(v|x)}{Q(\vec{y}|x)} \text{ if } v \in \mathcal{V}_x \text{ and } Q(\vec{y} \mid x) > 0, \\ 0 \text{ otherwise.} \end{cases}$$

Next, define $C \in \mathcal{C}$ according to

$$Q(c_j \mid v, x) = \begin{cases} \pi_j(x) \Big/ \left(\sum_{c_{\tilde{j}}:\, P_{\tilde{j}}(\cdot|x) = P(\cdot|v,x)} \pi_{\tilde{j}}(x) \right) & \text{if } v \in \mathcal{V}_x \text{ and } P_j(\cdot \mid c, x) = P(\cdot \mid v, x) \\ 0 \text{ otherwise.} \end{cases}$$

Together, this defines the random vector $(X, Y^*, V, C) \sim Q$, whose joint distribution is defined as

$$Q(x, \vec{y}, v, c) = P(x)Q(\vec{y} \mid x)Q(v \mid \vec{y}, x)Q(c \mid v, x).$$

We now check that this construction satisfies information set, expected utility maximization and data consistency. First, information set is satisfied since $Q(c, \vec{y} \mid x, v) = Q(\vec{y} \mid x, v)Q(c \mid x, v)$ by construction. Next, for any $x \in \mathcal{X}$ and $c_j \in \mathcal{C}_x$, define $v_{j,x} \in \mathcal{V}_x$ to be the label satisfying $P_j(\cdot \mid x) = P(\cdot \mid v, x)$. For $P(c_j, \vec{y} \mid w, x) > 0$, observe that

$$\begin{aligned} &P(c_j, \vec{y} \mid x) = \widetilde{P}_j(\vec{y} \mid x)\pi_j(x) = \\ &Q(\vec{y} \mid x) \frac{Q(\vec{y} \mid v_{j,x}, x) \sum_{\tilde{j}:\, P_{\tilde{j}}(\cdot|x) = P(\cdot|v_{j,x},x)} \pi_{\tilde{j}}(x)}{Q(\vec{y} \mid x)} \frac{\pi_j(x)}{\sum_{\tilde{j}:\, P_{\tilde{j}}(\cdot|x) = P(\cdot|v_{j,x},x)} \pi_{\tilde{j}}(x)} = \\ &Q(\vec{y} \mid x)Q(v_{j,x} \mid \vec{y}, x)Q(c \mid v_{j,x}, x) = \\ &\sum_{v \in \mathcal{V}} Q(\vec{y} \mid x)Q(v \mid \vec{y}, x)Q(c \mid v, x) = \sum_{v \in \mathcal{V}} Q_{V,C,\vec{Y}}(v, c, \vec{y} \mid x) = Q_{C,\vec{Y}}(c, \vec{y} \mid x). \end{aligned}$$

Moreover, whenever $P_{C,\vec{Y}}(c, \vec{y} \mid x) = 0$, $Q(y^* \mid v_{j,x}, x)Q(c \mid v_{j,x}, x) = 0$. Therefore, data consistency holds. Finally, by construction, for $Q(C = c_j \mid V = v_{j,x}, X = x) > 0$,

$$Q(\vec{Y} = \vec{y} \mid V = v_{j,x}, X = x) = \frac{Q(V = v_{j,x} \mid \vec{Y} = \vec{y}, X = x)Q(\vec{Y} = \vec{y} \mid X = x)}{Q(V = v_{j,x} \mid X = x)} = \widetilde{P}_j(\vec{y} \mid x).$$

Expected utility maximization is therefore satisfied since the inequalities in Lemma B.1 were assumed to hold and data consistency holds.

Lemma B.1 implies Theorem B.1: Define the joint distribution Q as $Q(x, c_j, \vec{y}) = \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)P(x)$. Rewrite the condition in Lemma B.1 as: for all $c_j \in \{c_1, \ldots, c_J\}$ and $c' \neq c_j$,

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x)u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x_I)u(c', \vec{y}; x_I).$$

Notice that if $\pi_j(x) = 0$, then $Q(c_j, \vec{y} \mid x) = 0$. The inequalities involving $c \in \mathcal{C}$ with $\pi_c(x) = 0$ are therefore satisfied. Next, inequalities involving $c_j \in \{c_1, \ldots, c_J\}$ with $\pi_j(x) > 0$ can be equivalently rewritten as

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q_j(\vec{y} \mid x)u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q_j(\vec{y} \mid x_I)u(c', \vec{y}; x_I).$$

The statement of Theorem B.1 follows by noticing that

$$\sum_{\vec{y}\in\mathcal{Y}^J} Q_j(\vec{y} \mid x)u(c_j, \vec{y}; x_I) = \mathbb{E}_Q\left[u(c_j, \vec{Y}; x_I) \mid C = c_j, X = x\right],$$
$$\sum_{\vec{y}\in\mathcal{Y}^J} Q_j(\vec{y} \mid x)U(c', \vec{y}; x_I) = \mathbb{E}_Q\left[u(c', \vec{Y}; x_I) \mid C = c_j, X = x\right].$$

□

B.5.2 Proof of Corollary B.1

Proof. This follows from Theorem B.1. For all $x \in \mathcal{X}$ with $\pi_1(x) > 0$, the inequalities require $\sum_{k=1}^K \mathbb{E}[Y_k(1) \mid C = 1, X = x] - u_{0,k}(x_I) \geq \sum_{k=1}^K \mathbb{E}[Y_k(0) \mid C = 1, X = x]$. For all $x \in \mathcal{X}$ with $\pi_0(x) > 0$, the inequalities require $\sum_{k=1}^K \mathbb{E}[Y_k(0) \mid C = 0, X = x] \geq \sum_{k=1}^K \mathbb{E}[Y_k(1) \mid C = 0, X = x] - u_{0,k}(x_I)$. Re-arranging delivers that the decision maker's choices are consistent with expected utility maximization at a linear utility function if and only if there exists $\widetilde{P}_1(\cdot \mid x) \in \mathcal{B}_{1,x}$ and $\widetilde{P}_0(\cdot \mid x) \in \mathcal{B}_{0,x}$ satisfying

$$\mu_0(x_I, x_E) \leq \sum_{k=1}^K |u_{0,k}(x_I)| \text{ whenever } \pi_0(x) > 0$$
$$\sum_{k=1}^K |u_{0,k}(x_I)| \leq \mu_1(x_I, x_E) \text{ whenever } \pi_1(x) > 0.$$

Taking the maximum of the upper bound over $\widetilde{P}_1(\cdot \mid x) \in \mathcal{B}_{1,x}$ and the minimum of the lower bound over $\widetilde{P}_0(\cdot \mid x) \in \mathcal{B}_{1,x}$ then yields the result. □

B.5.3 Proof of Theorem B.2

Proof. The proof follows the same strategy as Theorem B.1. I prove the following Lemma, and then show that it implies Theorem B.2.

Lemma B.2. *The decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a utility function $u \in \mathcal{U}$, $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ for each $c_j \in \{c_1, \ldots, c_J\}$ and $x \in \mathcal{X}$ satisfying*

$$\sum_{\vec{y}\in\mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y}\in\mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)u(c', \vec{y}; x_I) - \pi_j(x)\epsilon(x)$$

for all $x \in \mathcal{X}$, $c \in \{c_1, \ldots, c_J\}$, $c' \neq c_j$.

Proof of Lemma B.2: Necessity Suppose the decision maker's choices are consistent with approximate expected utility maximization behavior at some $u \in \mathcal{U}$, $(X, V, C, \vec{Y}) \sim Q$, and $\epsilon = \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$. By the same argument in the necessity direction for Lemma B.1, it is without loss of generality to assume $V \in \mathcal{V}$ has finite support.

For each $c_j \in \{c_1, \ldots, c_J\}$, $(x, v) \in \mathcal{X} \times \mathcal{V}$

$$Q(c_j \mid x, v) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c_j, \vec{y}; x_I) \right) \geq Q(c_j \mid x, v) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c', \vec{y}; x_I) \right) - Q(c_j \mid x, v) \epsilon(x)$$

holds for all $c_j \neq c'$. If $Q(c_j \mid x, v) = 0$, this holds trivially. If $Q(c_j \mid x, v) > 0$, this holds through the approximate expected utility maximization condition. Multiply both sides by $Q(v \mid x)$ to arrive at

$$Q(c_j \mid x, v) Q(v \mid x) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c_j, \vec{y}; x_I) \right) \geq$$

$$Q(c_j \mid x, v) Q(v \mid x) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c', \vec{y}; x_I) \right) - Q(c_j \mid x, v) Q(v \mid x) \epsilon(x).$$

Next, use information set to write $Q(c_j, \vec{y} \mid x, v) = Q(\vec{y} \mid x, v) Q(c_j \mid x, v)$ and arrive at

$$Q(v \mid x) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x, v) u(c_j, \vec{y}; x_I) \right) \geq Q(v \mid x) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x, v) u(c', \vec{y}; x_I) \right) - Q(c_j, v \mid x) \epsilon(x)$$

Finally, we use $Q(c_j, \vec{y}, v \mid x) = Q(c_j, \vec{y} \mid x, v) Q(v \mid x)$ and further sum over $v \in \mathcal{V}$ to arrive at

$$\sum_{\vec{y} \in \mathcal{Y}^J} \left(\sum_{v \in \mathcal{V}} Q(v, c_j, \vec{y} \mid x) \right) u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} \left(\sum_{v \in \mathcal{V}} Q(v, c_j, \vec{y} \mid x) \right) u(c', \vec{y}; x_I) - \sum_{v \in \mathcal{V}} Q(c_j, v \mid x) \epsilon(x),$$

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x) u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x) u(c', \vec{y}; x_I) - Q(c_j \mid x) \epsilon(x)$$

The inequalities in Lemma B.1 then follow from an application of data consistency.

Proof of Lemma B.2: Sufficiency Sufficiency follows by the same construction of the joint distribution $(X, V, Y^*, C) \sim Q$ as given in the sufficiency direction for Lemma B.1.

Lemma B.2 implies Theorem B.2 Define Q as $Q(x, c_j, \vec{y}) = \widetilde{P}_j(\vec{y} \mid x) \pi_j(x) P(x)$. Rewrite the condition in Lemma B.2 as: for all $c_j \in \{c_1, \ldots, c_J\}$ and $c' \neq c_j$,

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x) u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x_I) u(c', \vec{y}; x_I) - Q(c_j \mid x) \epsilon(x).$$

Notice that if $\pi_j(x) = 0$, then $Q(c_j, \vec{y} \mid x) = 0$ and $Q(c_j \mid x) = 0$. The inequalities involving $c \in \mathcal{C}$ with $\pi_c(x) = 0$ are therefore satisfied. The inequalities involving $c_j \in \{c_1, \ldots, c_J\}$ with

$\pi_j(x) > 0$ can be equivalently rewritten as

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q_j(\vec{y} \mid x) u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q_j(\vec{y} \mid x_I) u(c', \vec{y}; x_I) - \epsilon(x).$$

The statement of Theorem B.2 follows by noticing that

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q_j(\vec{y} \mid x) u(c_j, \vec{y}; x_I) = \mathbb{E}_Q \left[u(c_j, \vec{Y}; x_I) \mid C = c_j, X = x \right],$$
$$\sum_{\vec{y} \in \mathcal{Y}^J} Q_j(\vec{y} \mid x) u(c', \vec{y}; x_I) = \mathbb{E}_Q \left[u(c', \vec{Y}; x_I) \mid C = c_j, X = x \right].$$

□

B.5.4 Proof of Corollary B.3

Proof. I apply Theorem B.2 to a binary choice, treatment assignment problem over the class of linear utility functions. For all $(x_I, x_E) \in \mathcal{X}_I \times \mathcal{X}_E$ with $\pi_0(x_I, x_E) > 0$, Theorem B.2 requires $\underline{\mu}_0(x_I, x_E) - \epsilon(x_I, x_E) \leq \sum_{k=1}^K |u_{0,k}(x_I)|$. For all $(x_I, x_E) \in \mathcal{X}_I \times \mathcal{X}_E$ with $\pi_1(x_I, x_E) > 0$, Theorem B.2 requires $\sum_{k=1}^K |u_{0,k}(x_I)| \leq \overline{\mu}_1(x_I, x_E) + \epsilon(x_I, x_E)$. Putting these together, it follows that the decision maker's choices approximately maximize expected utility at $\epsilon = \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ if and only if, for all $x_I \in \mathcal{X}_I$,

$$\max_{x_E \in \mathcal{X}_E} \left\{ \underline{\mu}_0(x_I, x_E) - \epsilon(x_I, x_E) \right\} \leq \min_{x'_E \in \mathcal{X}_E} \left\{ \overline{\mu}_1(x_I, x_E) + \epsilon(x_I, x_E) \right\}.$$

This is equivalent to, for all $x_I \in \mathcal{X}_I$,

$$\underline{\mu}_0(x_I, x_E) - \overline{\mu}_1(x_I, x'_E) - \epsilon(x_I, x_E) - \epsilon(x_I, x'_E) \leq 0 \text{ for all } x_E, x'_E \in \mathcal{X}_E.$$

□

B.5.5 Proof of Theorem B.3

Proof. To prove this result, I first establish the following lemma, and then show Theorem B.3 follows as a consequence.

Lemma B.3. *Assume $\widetilde{P}(\cdot \mid x) > 0$ for all $\widetilde{P}(\cdot \mid x) \in \mathcal{H}(P(\cdot \mid x); \mathcal{B}_x)$ and all $x \in \mathcal{X}$. The decision maker's choices are consistent with expected utility maximization behavior at inaccurate beliefs if and only if there exists a utility function $u \in \mathcal{U}$, prior beliefs $Q(\cdot \mid x) \in \Delta(\mathcal{Y}^J)$ for all $x \in \mathcal{X}$, $\widetilde{P}_j(\cdot \mid x)$ for $j = 1, \ldots, J$ and all $x \in \mathcal{X}$ satisfying, for all $c_j \in \{c_1, \ldots, c_J\}$ and $c' \neq c_j$,*

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x) \widetilde{P}(c_j \mid \vec{y}, x) u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x) \widetilde{P}(c_j \mid \vec{y}, x) u(c', \vec{y}; x_I),$$

where $\widetilde{P}(c_j \mid \vec{y}, x) = \frac{\widetilde{P}_j(\vec{y} \mid x) \pi_j(x)}{\widetilde{P}(\vec{y} \mid x)}$ and $\widetilde{P}(\vec{y} \mid x) = \sum_{j=1}^J \widetilde{P}_j(\vec{y} \mid x) \pi_j(x)$.

Proof of Lemma B.3: Necessity First, by an analogous argument as given in the proof of necessity for Lemma B.1, it is without loss to assume $V \in \mathcal{V}$ has finite support. Second, following the same steps as the proof of necessity for Lemma B.1, I arrive at

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x) u(c, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x) u(c', \vec{y}; x_I).$$

We then observe $Q(c, \vec{y} \mid x) = Q(c \mid \vec{y}, x) Q(\vec{y} \mid x) = \widetilde{P}(c \mid \vec{y}, x) Q(\vec{y} \mid x)$, where the last equality follows via Data Consistency with Inaccurate Beliefs.

Proof of Lemma B.3: Sufficiency To show sufficiency, suppose that the conditions in Lemma B.3 are satisfied at some $\widetilde{P}_j(\cdot \mid x) \in \mathcal{B}_{j,x}$ for $c \in \{c_1, \ldots, c_J\}$, $x \in \mathcal{X}$ and some $Q(\cdot \mid x) \in \Delta(\mathcal{Y}^J)$ for all $x \in \mathcal{X}$.

Define the joint distribution $(X, C, \vec{Y}) \sim \widetilde{P}$ according to $\widetilde{P}(x, c, \vec{y}) = \widetilde{P}(c \mid \vec{y}, x) Q(\vec{y} \mid x) P(X = x)$, where $\widetilde{P}(c \mid \vec{y}, x)$ is defined in the statement of the Lemma. Given the inequalities in the Lemma, we construct a joint distribution $(X, V, C, \vec{Y}) \sim Q$ to satisfy information set, expected utility maximization behavior and data consistency with inaccurate beliefs for the constructed joint distribution $(X, C, \vec{Y}) \sim \widetilde{P}$ following the same sufficiency argument as given in Lemma B.1.

Let $\mathcal{C} = \{c_1, \ldots, c_J\}$ and $v \in \mathcal{V} := \{1, \ldots, 2^J\}$ index all possible subsets in the power set $2^{\mathcal{C}}$ Define $\widetilde{\pi}_j(x)$ to be the probability of $C = c_j$ given $X = x$ and $\widetilde{P}_j(\vec{y} \mid x)$ to be the conditional potential outcome probability given $C = c_j, X = x$ under the constructed joint distribution $(X, C, \vec{Y}) \sim \widetilde{P}$ in the statement of the Lemma.

For each $x \in \mathcal{X}$, define $\mathcal{C}_x := \{c_j \colon \widetilde{\pi}_j(x) > 0\} \subseteq \mathcal{C}$ to be the set of choices selected with positive probability, and partition $\mathcal{C}_x$ into subsets that have identical constructed conditional potential outcome probabilities. There are $\bar{V}_x \leq |\mathcal{C}_x|$ such subsets. Associate each subset in this partition with its associated index $v \in \mathcal{V}$ and denote the possible values as $\mathcal{V}_x$. Denote the choice-dependent outcome probability associated with the subset labelled v by $\widetilde{P}(\cdot \mid v, x) \in \Delta(\mathcal{Y}^J)$.

Define $V \in \mathcal{V}$ according to

$$Q(V = v \mid x) = \sum_{c_j \colon \widetilde{P}_j(\cdot|x) = \widetilde{P}(\cdot|v,x)} \widetilde{\pi}_j(x) \text{ if } v \in \mathcal{V}_x,$$

$$Q(V = v \mid \vec{y}, x) = \begin{cases} \frac{\widetilde{P}(\vec{y}|v,x) Q(V=v|x)}{Q(\vec{y}|x)} & \text{if } v \in \mathcal{V}_x \text{ and } Q(\vec{y} \mid x) > 0, \\ 0 \text{ otherwise.} \end{cases}$$

Next, define the random variable $C \in \mathcal{C}$ according to

$$Q(C = c_j \mid v, x) = \begin{cases} \frac{\widetilde{\pi}_j(x)}{\sum_{\tilde{j} \colon \widetilde{P}_{\tilde{j}}(\cdot|x) = \widetilde{P}(\cdot|v,x)} \widetilde{\pi}_{\tilde{j}}(x)} & \text{if } v \in \mathcal{V}_x \text{ and } \widetilde{P}_j(\cdot \mid x) = \widetilde{P}(\cdot \mid v, x) \\ 0 \text{ otherwise.} \end{cases}$$

Together, this defines the random vector $(X, \vec{Y}, V, C) \sim Q$, whose joint distribution is defined

as

$$Q(x, \vec{y}, v, c) = P(x)Q(\vec{y} \mid x)Q(v \mid \vec{y}, x)Q(c \mid v, x).$$

We check that this construction satisfies information set, expected utility maximization and data consistency. First, information set is satisfied since $Q(c, \vec{y} \mid x, v) = Q(\vec{y} \mid x, v)Q(c \mid x, v)$ by construction. Next, for any $x \in \mathcal{X}$ and $c_j \in \mathcal{C}_x$, define $v_{j,x} \in \mathcal{V}_x$ to be the label satisfying $\widetilde{P}_j(\cdot \mid x) = \widetilde{P}(\cdot \mid v, x)$. For $\widetilde{P}(c_j, \vec{y} \mid x) > 0$, observe that

$$\begin{aligned}
&\widetilde{P}(c_j, \vec{y} \mid x) = \widetilde{P}_j(\vec{y} \mid x)\widetilde{\pi}_j(x) = \\
&Q(\vec{y} \mid x)\frac{Q(\vec{y} \mid v_{j,x}, x)\sum_{\left\{\tilde{j}:\ \substack{\widetilde{P}_{\tilde{j}}(\cdot|x)=\\ \widetilde{P}(\cdot|v,x)}\right\}}\widetilde{\pi}_{\tilde{j}}(x)}{Q_{\vec{Y}}(\vec{y} \mid x)}\frac{\widetilde{\pi}_j(x)}{\sum_{\left\{\tilde{j}:\ \substack{\widetilde{P}_{\tilde{j}}(\cdot|x)=\\ \widetilde{P}(\cdot|v,x)}\right\}}\widetilde{\pi}_{\tilde{j}}(x)} = \\
&Q(\vec{y} \mid x)Q(v_{j,x} \mid \vec{y}, x)Q(c \mid v_{j,x}, x) = \\
&\sum_{v \in \mathcal{V}} Q(\vec{y} \mid x)Q(v \mid \vec{y}, x)Q(c \mid v, x) = \sum_{v \in \mathcal{V}} Q(v, c, \vec{y} \mid x).
\end{aligned}$$

Moreover, whenever $\widetilde{P}(c, \vec{y} \mid x) = 0$, $Q(\vec{y} \mid v_{j,x}, x)Q(c \mid v_{j,x}, x) = 0$. Since $\widetilde{P}(c, \vec{y} \mid x) = \widetilde{\pi}(c \mid \vec{y}, x)Q(\vec{y} \mid x)$ by construction, $(X, V, C, \vec{Y}) \sim Q$ satisfies data consistency at inaccurate beliefs (Definition B.4). Finally, for $Q(c_j \mid V = v_{j,x}, X = x) > 0$,

$$Q(\vec{Y} = \vec{y} \mid V = v_{j,x}, X = x) = \frac{Q(V = v_{j,x} \mid \vec{Y} = \vec{y}, X = x)Q(\vec{Y} = \vec{y} \mid X = x)}{Q(V = v_{j,x} \mid X = x)} = \widetilde{P}_j(\vec{y} \mid X = x)$$

and $\widetilde{\pi}_j(x) > 0$. Therefore, using data consistency at inaccurate beliefs and the inequalities in Lemma B.3, we have that

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid v, x)u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid v, x)u(c', \vec{y}; x_I),$$

which follows from the fact that $Q_{\vec{Y}}(\vec{y} \mid x)\widetilde{P}(c_j \mid \vec{y}, x) = Q(c_j, \vec{y} \mid x)$ and the construction of $\widetilde{P}$, and $Q(\vec{Y} = \vec{y} \mid V = v_{j,x}, X = x) = \widetilde{P}_j(\vec{y} \mid x)$ as just shown. Therefore, expected utility maximization is also satisfied.

Rewrite inequalities in Lemma B.3 in terms of weights: Define $\widetilde{P}$ as in the statement of the Theorem. Rewrite the condition in Lemma B.3 as: for all $c_j \in \{c_1, \ldots, c_J\}$ and $\tilde{c} \neq c_j$,

$$\sum_{\vec{y} \in \mathcal{Y}^J} \frac{Q(\vec{y} \mid x)}{\widetilde{P}(\vec{y} \mid x)}\widetilde{P}(c, \vec{y} \mid x)u(c, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} \frac{Q(\vec{y} \mid x)}{\widetilde{P}(\vec{y} \mid x)}\widetilde{P}(c, \vec{y} \mid x)u(c', \vec{y}; x_I).$$

Notice that if $\pi_j(w, x) = 0$, then $\widetilde{P}(c_j, \vec{y} \mid x) = 0$. Therefore, the inequalities involving $c_j \in \{c_1, \ldots, c_J\}$ with $\pi_j(x) = 0$ are trivially satisfied. The inequalities involving $c \in \{c_1, \ldots, c_J\}$

with $\pi_j(x) > 0$ can be equivalently rewritten as

$$\sum_{\vec{y}\in\mathcal{Y}^J} \frac{Q(\vec{y} \mid x)}{\widetilde{P}(\vec{y} \mid x)} \widetilde{P}_j(\vec{y} \mid x) u(c, \vec{y}; x_I) \geq \sum_{\vec{y}\in\mathcal{Y}^J} \frac{Q(\vec{y} \mid x)}{\widetilde{P}(\vec{y} \mid x)} \widetilde{P}_j(\vec{y} \mid x) u(c', \vec{y}; x_I).$$

The result follows by noticing $\sum_{\vec{y}\in\mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid x) \frac{Q(\vec{y}|x)}{\widetilde{P}(\vec{y}|x)} u(c, \vec{y}; x_I) = \mathbb{E}_{\widetilde{P}}\left[\frac{Q(\vec{y}|x)}{\widetilde{P}(\vec{y}|x)} u(c, \vec{y}; x_I)\right]$ and defining the weights as $\omega(\vec{y}; x) = \frac{Q(\vec{y}|x)}{\widetilde{P}(\vec{y}|x)}$. □

C Additional results for the econometric framework

C.1 Quasi-randomly assigned instrumental variable

In this section, I modify Assumption 3.2 to only require that the instrument be quasi-randomly assigned conditional on some additional finite support characteristics $t \in \mathcal{T}$. Consider the joint distribution $(X, T, Z, C, Y^*) \sim P(\cdot)$ and assume that it satisfies the following assumption.

Assumption C.1 (Quasi-Random Instrument)**.** The joint distribution $(X, T, Z, C, Y^*) \sim P(\cdot)$ satisfies $(X, Y^*) \perp\!\!\!\perp Z \mid T$ and there exists some $\delta > 0$ such that $P(x, t, z) := P(X = x, T = t, Z = z) \geq \delta$ for all $(x, t, z) \in \mathcal{X} \times \mathcal{T} \times \mathcal{Z}$.

In the empirical application to the New York City pretrial release system, judges are quasi-randomly assigned to cases within court-by-time strata $T \in \mathcal{T}$.

Under Assumption 3.1 and Assumption C.1, researchers can derive bounds on the unobservable conditional outcome probabilities $\mu_0(x, z)$, as needed to apply the characterization results for expected utility maximization at accurate beliefs provided in Section 3.1 of the main text. To state this result, let $\mu(x, z) := \mathbb{E}[\bar{Y}^* \mid X = x, Z = z]$, $\mu(x, t, z) := \mathbb{E}[\bar{Y}^* \mid X = x, Z = z, T = t]$, $\mu_c(x, t, z) := \mathbb{E}[\bar{Y}^* \mid C = c, X = x, Z = z, T = t]$ for $c \in \{0, 1\}$, $\pi_c(x, t, z) := P(C = c \mid X = x, T = t, Z = z)$ for $c \in \{0, 1\}$, and $P(t \mid x, z) := P(T = t \mid X = x, Z = z)$.

Proposition C.1. *Suppose Assumption 3.1 and Assumption C.1 hold. For any $(x, z) \in \mathcal{X} \times \mathcal{Z}$ with $\pi_0(x, z) > 0$, the identified set for $\mu_0(x, z)$ is the interval $[\underline{\mu}'_0(x, z), \overline{\mu}'_0(x, z)]$, where*

$$\underline{\mu}'_0(x, z) = \max\left\{\frac{\underline{\mu}'(x, z) - \mu_1(x, z)\pi_1(x, z)}{\pi_0(x, z)}, 0\right\}, \quad \overline{\mu}'_0(x, z) = \min\left\{\frac{\overline{\mu}'(x, z) - \mu_1(x, z)\pi_1(x, z)}{\pi_0(x, z)}, 1\right\},$$

$$\underline{\mu}'(x, z) = \mathbb{E}[\max_{\tilde{z}\in\mathcal{Z}}\{\mu_1(x, \tilde{z}, T)\pi_1(x, \tilde{z}, T)\} \mid X = x, Z = z], \text{ and}$$

$$\overline{\mu}'(x, z) = \mathbb{E}[\min_{\tilde{z}\in\mathcal{Z}}\{\mu_1(x, \tilde{z}, T)\pi_1(X, \tilde{z}, T) + K\pi_0(X, \tilde{z}, T)\} \mid X = x, Z = z].$$

Proof. Under Assumption C.1, $\mu(x, t) = \mu(x, t, z) = \mu(x, t, \tilde{z})$ for any $x \in \mathcal{X}$, $t \in \mathcal{T}$, and $z, \tilde{z} \in \mathcal{Z}$. By the same reasoning as in the proof of Proposition 3.1, $\mu(x, t, z)$ is bounded by

$$\mu_1(x, t, \tilde{z})\pi_1(x, t, \tilde{z}) \leq \mu(x, t, z) \leq K\pi_0(x, t, \tilde{z}) + \mu_1(x, t, \tilde{z})\pi_1(x, t, \tilde{z}).$$

Therefore, for any $x \in \mathcal{X}$, $z \in \mathcal{Z}$, $\mu(x,z)$ satisfies

$$\mathbb{E}\left[\max_{\tilde{z}\in\mathcal{Z}}\{\mu_1(x,\tilde{z},T)\pi_1(x,\tilde{z},T)\} \mid X=x, Z=z\right] \leq \mu(x,z),$$
$$\mu(x,z) \leq \mathbb{E}\left[\min_{\tilde{z}\in\mathcal{Z}}\{\mu_1(x,\tilde{z},T)\pi_1(X,\tilde{z},T) + K\pi_0(X,\tilde{z},T)\} \mid X=x, Z=z\right].$$

The result then follows immediately via iterated expectations. □

Under Assumption 3.1 and Assumption C.1, the researcher can also derive valid bounds at any particular instrument value by the same logic. In particular, for any $\tilde{z} \in \mathcal{Z}$, $\mu_0(x,z)$ is bounded below and above respectively by

$$\underline{\mu}'_{0,\tilde{z}}(x,z) = \max\left\{\frac{\underline{\mu}'_{\tilde{z}}(x,z) - \mu_1(x,z)\pi_1(x,z)}{\pi_0(x,z)}, 0\right\},\ \overline{\mu}'_{0,\tilde{z}}(x,z) = \min\left\{\frac{\overline{\mu}'_{\tilde{z}}(x,z) - \mu_1(x,z)\pi_1(x,z)}{\pi_0(x,z)}, 1\right\},$$

where $\underline{\mu}'_{\tilde{z}}(x,z) = \mathbb{E}[\{\mu_1(x,\tilde{z},T)\pi_1(x,\tilde{z},T)\} \mid X=x, Z=z]$ and $\overline{\mu}'_{\tilde{z}}(x,z) = \mathbb{E}[\mu_1(x,\tilde{z},T)\pi_1(X,\tilde{z},T) + K\pi_0(X,\tilde{z},T) \mid X=x, Z=z]$.

Assumption C.1 imposed that $P(x,t,z) \geq \delta$ for all $(x,t,z) \in \mathcal{X}\times\mathcal{T}\times\mathcal{Z}$ and some $\delta > 0$. This implies $P(Z=z \mid X=x, T=t) > 0$ for all values $(x,t,z) \in \mathcal{X}\times\mathcal{T}\times\mathcal{Z}$, and so instrument is assumed to satisfy strict overlap conditional on the characteristics X and the additional characteristic T. Of course, strict overlap may be violated in particular empirical applications, and the bounds can be suitably extended.

Proposition C.2. *Suppose Assumption 3.1 is satisfied, and the joint distribution $(X,T,Z,C,Y^*) \sim P(\cdot)$ satisfies $(X,Y^*) \perp\!\!\!\perp Z \mid T$. For any $(x,z) \in \mathcal{X}\times\mathcal{Z}$ with $\pi_0(x,z) > 0$, $\mu_0(x,z)$ is bounded below and above respectively by*

$$\underline{\mu}''_0(x,z) = \max\left\{\frac{\underline{\mu}''(x,z) - \mu_1(x,z)\pi_1(x,z)}{\pi_0(x,z)}, 0\right\} \text{ and } \overline{\mu}''_0(x,z) = \min\left\{\frac{\overline{\mu}''(x,z) - \mu_1(x,z)\pi_1(x,z)}{\pi_0(x,z)}, 1\right\},$$

where $\mathcal{Z}(x,t) = \{z\colon P(Z=z \mid X=x, T=t) > 0\}$, $\underline{\mu}''(x,z) = \mathbb{E}[\max_{\tilde{z}\in\mathcal{Z}(x,T)}\{\mu_1(x,\tilde{z},T)\pi_1(x,\tilde{z},T)\} \mid X=x, Z=z]$ and $\overline{\mu}''(x,z) = \mathbb{E}[\min_{\tilde{z}\in\mathcal{Z}(x,T)}\{\mu_1(x,\tilde{z},T)\pi_1(X,\tilde{z},T) + K\pi_0(X,\tilde{z},T)\} \mid X=x, Z=z]$.

Proof. The proof follows the same argument as Proposition C.1. □

C.2 Translating expected utility costs into ex-post errors

Section 4.1.1 of the main text showed that the total expected utility cost $\underline{\mathcal{E}}$ of systematic prediction mistakes to the decision maker can be characterized as the optimal value of a linear program. For a scalar outcome $Y^* = Y_1^*$, $\underline{\mathcal{E}}$ can be translated into an equivalent reduction in ex-post errors $\mathbb{E}[Y_1^* \cdot C]$ that would produce the same expected utility cost $\underline{\mathcal{E}}$ to the decision maker.

Assume $Y^* = Y_1^*$, and let $\bar{\epsilon}(x)$ denote an optimal solution to Equation (4). By the

definition of approximate expected utility maximization, $\underline{\mathcal{E}}$ is an upper bound on

$$\mathbb{E}[u(C^*(X,V),Y_1^*;X_I)-u(C,Y_1^*;X_I)] = \sum_{x_I\in\mathcal{X}_I}\left\{|u_{0,1}(x_I)|\Delta_{0,0}(x_I)+|u_{1,1}(x_I)|\Delta_{1,1}(x_I)\right\}P(x_I), \tag{17}$$

where $C^*(X,V)$ is the optimal choice at (X,V), $\Delta_{0,0}(x_I) = \mathbb{E}[(1-C)(1-Y_1^*) - (1-C^*(X,V))(1-Y_1^*) \mid X_I = x_I]$ is the reduction of ex-post errors that select $C=0$ when Y_1^* is small, and $\Delta_{1,1}(x_I) = \mathbb{E}[CY_1^* - C^*(X,V)Y_1^* \mid X_I = x_I]$ is the reduction of ex-post errors that select $C=1$ when Y_1^* is large. From the proof of Theorem 4.1, the identified set of linear utility functions at expected utility costs $\overline{\epsilon}(x)$ is satisfies, for all $x_I \in \mathcal{X}_I$,

$$\max_{\tilde{x}_E\in\mathcal{X}_E}\left\{\mu_1(x_I,\tilde{x}_E)-\overline{\epsilon}(x_I,\tilde{x}_E)\right\} \le |u_{0,1}(x_I)| \le \min_{\tilde{x}_E\in\mathcal{X}_E}\max_{\tilde{x}_E\in\mathcal{X}}\left\{\overline{\mu}_0(x_I,\tilde{x}_E)-\overline{\epsilon}(x_I,\tilde{x}_E)\right\}, \tag{18}$$

and define $|u_{0,1}(x_I)| = \max_{\tilde{x}_E\in\mathcal{X}_E}\left\{\mu_1(x_I,\tilde{x}_E)-\overline{\epsilon}(x_I,\tilde{x}_E)\right\}$, $|u_{1,1}(x_I)| = 1-|u_{0,1}(x_I)|$. At this candidate linear utility function, we can calculate the implied reduction in ex-post errors $\Delta_{1,1}(x_I)$ that are equivalent to $\underline{\mathcal{E}}$ in an expected utility sense by calculating

$$\begin{aligned}\max_{\Delta_{0,0}(x_I),\Delta_{1,1}(x_I)} & \sum_{x_I}\Delta_{1,1}(x_I)P(x_I) \\ \text{s.t. } & 0\le\Delta_{1,1}(x_I)\le\mathbb{E}[CY^*\mid X_I=x_I] \text{ for all } x_I\in\mathcal{X}_I, \\ & 0\le\Delta_{0,0}(x_I)\le\pi_0(x_I) \text{ for all } x_I\in\mathcal{X}_I, \\ & \sum_{x_I\in\mathcal{X}_I}\left\{|u_{0,1}(x_I)|\Delta_{0,0}(x_I)+|u_{1,1}(x_I)|\Delta_{1,1}(x_I)\right\}P(x_I)\le\underline{\mathcal{E}}.\end{aligned} \tag{19}$$

The first constraint imposes that the reduction in ex-post errors $\Delta_{1,1}(x_I)$ must be weakly positive, and can be no greater than the observed ex-post errors $\mathbb{E}[CY_1^* \mid X_I = x_I]$ at the decision maker's choices. The second constraint imposes that the reduction in ex-post errors $\Delta_{0,0}(x_I)$ must also be weakly positive, and can be no greater than the observed probability the decision maker selected $C=0$. The final constraint imposes that the implied expected utility of the change in ex-post errors must be consistent with the expected utility cost $\underline{\mathcal{E}}$.

The implied reduction in ex-post errors computed by Equation (19) is an accounting exercise to better summarize the magnitudes of the decision maker's implied prediction mistakes. At the conjectured utility function $u_{0,1}(\cdot), u_{1,1}(\cdot)$, it searches for the largest reduction in ex-post errors $\Delta_{0,0}(\cdot), \Delta_{1,1}(\cdot)$ that is consistent with the total expected utility cost $\underline{\mathcal{E}}$. The constraints do not impose that the ex-post errors must be achievable by behavior that is consistent with expected utility maximization at accurate beliefs. As a result, the optimal value of Equation (19) cannot be interpreted as a feasible reduction in ex-post errors $\mathbb{E}[Y_1^*\cdot d]$ that could be achieved by expected utility maximization at accurate beliefs. Furthermore, notice that if $\sum_{x_I\in\mathcal{X}_I}|u_{1,1}(x_I)|\mathbb{E}[Y_1^*\cdot C \mid X_I=x_I]P(x_I)\le\underline{\mathcal{E}}$, then the optimal value of Equation (19) is trivially equal to $\mathbb{E}[Y_1^*\cdot C]$.

C.3 Characterizing the share of systematic prediction mistakes

In Section 4.1.2, I defined the largest rationalizable subset $\overline{\mathcal{X}}$ of characteristics is then

$$\overline{\mathcal{X}} := \arg\max_{\tilde{\mathcal{X}} \subseteq \mathcal{X}} \sum_{x \in \tilde{\mathcal{X}}} P(x) \text{ s.t. } \tilde{\mathcal{X}} \text{ is rationalizable at accurate beliefs.} \tag{20}$$

The share of rationalizable decisions is $P(\overline{\mathcal{X}}) := \sum_{x \in \overline{\mathcal{X}}} P(x)$, and the *share of systematic prediction mistakes* in the decision maker's choices is therefore $1 - P(\overline{\mathcal{X}})$. Theorem 4.1 implies that the share of systematic prediction mistakes can be equivalently characterized by the following optimization program.

Theorem C.1. *The share of systematic prediction mistakes in the decision maker's choices satisfies* $1 - P(\overline{\mathcal{X}}) = \underline{\lambda}$*, where*

$$\begin{aligned} \underline{\lambda} := &\min_{\epsilon} \sum_{x \in \mathcal{X}} P(x) 1\{\epsilon(x) > 0\} \\ &\text{s.t. } \epsilon(x) \geq 0 \text{ for all } x \in \mathcal{X}, \\ &\quad \mu_1(x) - \overline{\mu}_0(x') - \epsilon(x) - \epsilon(x') \leq 0 \text{ for all pairs } x = (x_I, x_E), x' = (x_I, x'_E). \end{aligned} \tag{21}$$

Proof. To prove this result, I will show that $1 - \underline{\lambda} = P(\overline{\mathcal{X}})$. First, let $\epsilon^*(x)$ denote an optimal solution to the program defining $\underline{\lambda}$, meaning $\underline{\lambda} = \sum_{x \in \mathcal{X}} P(x) 1\{\epsilon^*(x) > 0\}$. Define $\mathcal{X}_R = \{x \in \mathcal{X} \colon \epsilon^*(x) = 0\}$, and observe that $\mathcal{X}_R$ is a rationalizable subset at accurate beliefs, since, for all pairs $(x_I, x_E), (x_I, x'_1) \in \mathcal{X}_R$, $\mu_1(x_I, x_E) - \overline{\mu}_0(x_I, x'_E) \leq 0$ by construction. As a consequence, $P(\overline{\mathcal{X}}) \geq P(\mathcal{X}_R) = \sum_{x \in \mathcal{X}} P(x) 1\{\epsilon^*(x) = 0\} = 1 - \underline{\lambda}$.

Next, for each $x \in \overline{\mathcal{X}}$, define $\overline{\epsilon}(x) = 0$. For each $x = (x_I, x_E) \notin \overline{\mathcal{X}}$, define

$$\overline{\epsilon}_1(x) = \max_{x'_E} \{\mu_1(x) - \overline{\mu}_0(x_I, x'_E)\}, \text{ and } \overline{\epsilon}_2(x) = \max_{x'_E} \{\mu_1(x_I, x'_E) - \overline{\mu}_0(x)\},$$

and set $\overline{\epsilon}(x) = \max\{\overline{\epsilon}_1(x), \overline{\epsilon}_2(x)\}$. By construction, $\mu_1(x_I, x_E) - \overline{\mu}_0(x_I, x'_E) - \overline{\epsilon}(x_I, x_E) - \overline{\epsilon}(x_I, x'_E) \leq 0$ for all pairs $(x_I, x_E), (x_I, x'_E) \in \mathcal{X}$. $\overline{\epsilon} = \{\overline{\epsilon}(x) \colon x \in \mathcal{X}\}$ is therefore feasible in the program defining $\underline{\lambda}$, and so

$$\underline{\lambda} \leq \sum_{x \in \mathcal{X}} P(x) 1\{\overline{\epsilon}(x) > 0\}.$$

This implies $1 - \underline{\lambda} \geq 1 - \sum_{x \in \mathcal{X}} P(x) 1\{\overline{\epsilon}(x) > 0\} = \sum_{x \in \mathcal{X}} P(x) 1\{\overline{\epsilon}(x) = 0\} = P(\overline{\mathcal{X}})$. □

The share of systematic prediction mistakes in the decision maker's choices $\underline{\lambda}$ defined in (21) can be equivalently written as the optimal value of a mixed-integer linear program. This uses the standard "Big-M" method. Defining $M \geq 2K$ to be some large known constant, it

follows

$$\underline{\lambda} := \min_{\omega(x),\epsilon(x)} \sum_x P(x)\omega(x) \text{ s.t.}$$
$$\mu_1(x) - \overline{\mu}_0(x')\epsilon(x) - \epsilon(x') \leq 0 \text{ for all pairs } x = (x_I, x_E), x' = (x_I, x'_E),$$
$$0 \leq \epsilon(x) \leq M \cdot \omega(x),\ \omega(x) \in \{0,1\}$$

since $\max_{x,x' \in \mathcal{X}}\{\mu_1(x) - \overline{\mu}_0(x')\} \leq 2K$ because $Y_k^* \in [0,1]$ for all $k = 1, \ldots, K$.

C.4 Extension to incorporating monetary bail conditions

In Section 5, I defined the judge's choice to be a binary choice of whether to release or detain the defendant. In practice, judges in New York City choose what bail conditions and monetary amount to set for a defendant. Defendants may either be "released on recognizance" (i.e., released without bail conditions) or the judge may set bail conditions, in which case the defendant is only released if they can post the set bail amount. As a robustness exercise, I now define a judge's choice as whether or not to release the defendant on recognizance.

Let $C \in \{0,1\}$ denote whether the judge released the defendant "on recognizance" ($C = 1$). The outcome is $Y^* = (R^*, Y_1^*)$, where $R^* \in \{0,1\}$ denotes whether the defendant would satisfy the monetary bail condition set by the judge and $Y_1^* \in \{0,1\}$ is whether the defendant would fail to appear in court if released. Let $R \in \{0,1\}$ denote whether the defendant was released. The observed release satisfies $R = C + (1-C)R^*$, meaning the defendant is released if the judge selects release on recognizance or sets monetary bail conditions and the defendant satisfies them. I assume the judge payoffs only depend on whether a defendant is released and fails to appear in court or a defendant is detained and would not fail to appear in court. That is, I consider the set of utility functions $\mathcal{U}$ satisfying $u(c, r^*, y_1^*; x_I) = u(r, y_1^*; x_I)$, where $u(0,1;x_I) = 0, u(1,0;x_I) = 0$, $u(0,0;x_I) < 0, u(1,1;x_I) < 0$, and $|u(0,0;x_I) + u(1,1;x_I)| = 1$.

I apply Theorem B.1 to derive conditions under which the judge's choices are consistent with expected utility maximization at accurate beliefs about both failure to appear risk and the ability of defendant's to meet the bail conditions. For each $x_I \in \mathcal{X}_I$, define $\Pi_1(x_I) := \{x_E \in \mathcal{X}_E \colon \pi_1(x_I, x_E) > 0\}$ and $\Pi_0(x_I) := \{x_E \in \mathcal{X}_E \colon \pi_0(x_I, x_E) > 0\}$.

Proposition C.3. *Assume $P(Y_1^* = 1 \mid X = x) < 1$ for all $x \in \mathcal{X}$ with $\pi_1(x) > 0$ and $P(R = 0 \mid C = 0, X = x) > 0$ for all $x \in \mathcal{X}$ with $\pi_0(x) > 0$. The decision maker's choices are consistent with expected utility maximization behavior at some $u \in \mathcal{U}$ if and only if, for all $x_I \in \mathcal{X}_I$,*

$$\max_{x_E \in \Pi_1(x_I)} P(Y_1^* = 1 \mid C = 1, X = (x_I, x_E)) \leq \min_{x_E \in \Pi_0(x_I)} P(Y_1^* = 1 \mid R = 0, C = 0, X = (x_I, x_E)).$$

Proof. The inequalities in Theorem B.1 imply that the judge's choices are consistent with expected utility maximization behavior at accurate beliefs if and only if

(1) for all $x \in \mathcal{X}$ with $\pi_1(x) > 0$, $P(Y_1^* = 1 \mid C = 1, X = x) \leq -u(0,0;x_I)$.

(2) for all $x \in \mathcal{X}$ with $\pi_0(x) > 0$,

$$P(Y_1^* = 1, R = 1 \mid C = 0, X = x)u(1,1;x_I) + P(Y^* = 0, R = 0 \mid C = 0, X = x)u(0,0;x_I) \geq$$

$$P(Y_1^* = 1 \mid C = 0, X = x)u(1, 1; x_I).$$

The condition (2) may be re-arranged as

$$P(Y_1^* = 0, R = 0 \mid C = 0, X = x)u(0, 0; x_I) \geq P(Y_1^* = 1, R = 0 \mid C = 0, X = x)u(1, 1; x_I),$$

where $P(Y_1^* = 0, R = 0 \mid C = 0, X = x) = P(R = 0 \mid C = 0, X = x) - P(Y_1^* = 1, R = 0 \mid C = 0, X = x)$. Substituting and re-arranging then delivers

$$P(Y_1^* = 1, R = 0 \mid C = 0, X = x)\left(-u(0, 0; x_I) - u(1, 1; x_I)\right) \geq$$

$$-P(R = 0 \mid C = 0, X = x)u(0, 0; x_I).$$

The result follows. □

In Section 5.4.2 of the main text, I test whether the inequalities in Proposition C.3 are satisfied over the deciles of predicted failure to appear risk constructed in Section 5.2. I use the quasi-random assignment of judges to cases to construct bounds on the unobservable failure to appear rate among detained defendants that could not satisfy their monetary bail conditions. I now estimate the observed failure to appear rate only among defendants that were released on recognizance.

C.5 Expected social welfare: identification and inference

C.5.1 Expected social welfare under candidate decision rules

For a binary outcome $Y^* = Y_1^* \in \{0, 1\}$, consider a policymaker whose payoffs are summarized by the linear social welfare function $u_{1,1}^* y_1^* c + u_{0,1}^*(1 - y_1^*)(1 - c)$ as in Section 6 of the main text. As notation, let $\theta(x)$ denote expected social welfare at $x \in \mathcal{X}$ under a candidate decision rule $\tilde{\pi}(x)$ as given in (11), which can be rewritten as

$$\theta(x) = \ell(x; \tilde{\pi}, u^*)P(1 \mid x) + \beta(x; \tilde{\pi}, u^*) \tag{22}$$

for $\ell(x; \tilde{\pi}, u^*) := u_{1,1}^* \tilde{\pi}(x) - u_{0,1}^*(1 - \tilde{\pi}(x))$ and $\beta(x; \tilde{\pi}, u^*) := u_{0,1}^*(1 - \tilde{\pi}(x))$. Total expected social welfare then equals

$$\theta(\tilde{\pi}, u^*) = \beta(\tilde{\pi}, u^*) + \sum_{x \in \mathcal{X}} P(x)\ell(x; \tilde{\pi}, u^*)P(1 \mid x), \tag{23}$$

where $\beta(\tilde{\pi}, u^*) := \sum_{x \in \mathcal{X}} P(x)\beta(x; \tilde{\pi}, u^*)$. Since $P(1 \mid x)$ is partially identified, total expected social welfare is also partially identified and its sharp identified set of total expected welfare is an interval.

Proposition C.4. *Assume a binary outcome $Y^* = Y_1^* \in \{0, 1\}$. Consider a policymaker with a linear social welfare function $u_{0,1}^*, u_{1,1}^* < 0$ and a candidate decision rule $\tilde{\pi}(x)$. The sharp identified set of total expected social welfare, denoted $\mathcal{H}(\theta(\tilde{\pi}, u^*); \mathcal{B})$, is an interval with*

$\mathcal{H}(\theta(\tilde{\pi}, u^*); \mathcal{B}) = \left[\underline{\theta}(\tilde{\pi}, u^*), \overline{\theta}(\tilde{\pi}, u^*)\right]$, *where*

$$\underline{\theta}(\tilde{\pi}, u^*) = \beta(\tilde{\pi}, u^*) + \left\{ \min_{\left\{\widetilde{P}(\cdot|x):\ x\in\mathcal{X}\right\}} \sum_{x\in\mathcal{X}} P(x)\ell(x;\tilde{\pi},u^*)\widetilde{P}(1 \mid x) \ \ s.t.\ \ \widetilde{P}(\cdot \mid x) \in \mathcal{H}(P(\cdot \mid x); \mathcal{B}_{0,x}) \ \forall x \in \mathcal{X} \right\},$$

$$\overline{\theta}(\tilde{\pi}, u^*) = \beta(\tilde{\pi}, u^*) + \left\{ \max_{\left\{\widetilde{P}(\cdot|x):\ x\in\mathcal{X}\right\}} \sum_{x\in\mathcal{X}} P(x)\ell(x;\tilde{\pi},u^*)\widetilde{P}(1 \mid x) \ \ s.t.\ \ \widetilde{P}(\cdot \mid x) \in \mathcal{H}(P(\cdot \mid x); \mathcal{B}_{0,x}) \ \forall x \in \mathcal{X} \right\}.$$

For a binary outcome, the bounds $\mathcal{B}_x$ can be expressed as an interval with $[\underline{P}(1 \mid x), \overline{P}(1 \mid x)]$. For example, this is true if the bounds are constructed using an instrumental variable as discussed in the main text. In this case, Proposition C.4 implies that the sharp identified set of total expected social welfare under a candidate decision rule is characterized by the solution to two linear programs. Furthermore, provided the candidate decision rule and joint distribution of the characteristics X are known, testing the null hypothesis that total expected social welfare is equal to some candidate value is equivalent to testing a system of moment inequalities with nuisance parameters that enter linearly.

Proposition C.5. *Under the same set-up as Proposition C.4, conditional on the characteristics X, testing the null hypothesis $H_0\colon \theta(\tilde{\pi}, u^*) = \theta_0$ is equivalent to testing whether*

$$\exists \delta \in \mathbb{R}^{d_x - 1} \ s.t.\ \widetilde{A}_{(\cdot,1)}\left(\theta_0 - \ell^\intercal(\tilde{\pi}, u^*)P^{c=1,y_1^*=1} - \beta(\tilde{\pi}, u^*)\right) + \widetilde{A}_{(\cdot,-1)}\delta \leq \begin{pmatrix} -\underline{P}^{c=0,y_1^*=1} \\ \overline{P}^{c=0,y_1^*=1} \end{pmatrix},$$

where $d_x := |\mathcal{X}|$, $\ell(\tilde{\pi}, u^*)$ is the d_x-dimensional vector with elements $P(x)\ell(x;\tilde{\pi},u^*)$, $P^{c=1,y_1^*=1}$ is the d_x-dimensional vector of moments $P(C=1, Y_1^*=1 \mid X=x)$, $\underline{P}^{c=0,y_1^*=1}, \overline{P}^{c=0,y_1^*=1}$ are the d_x-dimensional vectors of lower and upper bounds on $P(C=0, Y_1^*=1 \mid X=x)$ respectively, and $\widetilde{A}$ is a known matrix.[6]

A confidence interval for total expected social welfare can then be constructed through test inversion. Testing procedures for moment inequalities with nuisance parameters are available for high-dimensional settings in Belloni, Bugni and Chernozhukov (2018). Andrews, Roth and Pakes (2023) and Cox and Shi (2022) develop inference procedures that exploit the additional linear structure and are valid in low-dimensional settings.

C.5.2 Expected social welfare under decision maker's observed choices

Consider again a policymaker with linear social welfare function $u_{0,1}^* < 0, u_{1,1}^* < 0$. Total expected social welfare under the decision maker's observed choices is given by

$$\begin{aligned}\theta^{DM}(u^*) =& u_{1,1}^* P(C=1, Y_1^*=1) + u_{0,1}^* P(C=0) \\ & - u_{0,1}^* \sum_{x\in\mathcal{X}} P(C=0, Y_1^*=1 \mid X=x)P(x).\end{aligned}$$

[6] For a matrix B, $B_{(\cdot,1)}$ refers to its first column and $B_{(\cdot,-1)}$ refers to all columns except its first column.

Since $P(C=0, Y_1^*=1 \mid X=x)$ is partially identified, total expected social welfare under the decision maker's observed choices is also partially identified and the sharp identified set is again an interval.

Proposition C.6. *Under the same set-up as Proposition C.4, the sharp identified set of total expected social welfare under the decision maker's observed choices, denoted $\mathcal{H}(\theta^{DM}(u^*); \mathcal{B})$, is an interval with $\mathcal{H}(\theta^{DM}(u^*); \mathcal{B}) = \left[\underline{\theta}^{DM}(u^*), \overline{\theta}^{DM}(u^*)\right]$, where*

$$\underline{\theta}^{DM}(u^*) = u^*_{1,1} P(C=1, Y_1^*=1) + u^*_{0,1} P(C=0) - u^*_{0,1} \overline{P}(C=0, Y_1^*=1)$$
$$\overline{\theta}^{DM}(u^*) = u^*_{1,1} P(C=1, Y_1^*=1) + u^*_{0,1} P(C=0) - u^*_{0,1} \underline{P}(C=0, Y_1^*=1),$$

where

$$\overline{P}(C=0, Y_1^*=1) = \max_{\left\{\widetilde{P}(C=0, Y_1^*=1 \mid X=x):\ x \in \mathcal{X}\right\}} \sum_{x \in \mathcal{X}} P(x) \widetilde{P}(C=0, Y_1^*=1 \mid X=x)$$
$$s.t.\ \widetilde{P}(C=0, Y_1^*=1 \mid X=x) \in \mathcal{H}(P(C=0, Y_1^*=1 \mid X=x); \mathcal{B}_{0,x})\ \forall x \in \mathcal{X}$$

and $\underline{P}(C=0, Y_1^*=1)$ is the optimal value of the analogous minimization problem.

The bounds $\mathcal{B}_x$ for a binary outcome are an interval, and so Proposition C.4 implies that the sharp identified set of total expected social welfare under a candidate decision rule is characterized by the solution to two linear programs.

Provided the joint distribution of the characteristics X are known, then testing the null hypothesis that total expected social welfare is equal to some candidate value is equivalent to testing a system of moment inequalities with a large number of nuisance parameters that enter the moments linearly.

Proposition C.7. *Under the same set-up as Proposition C.4, conditional on the characteristics X, testing the null hypothesis $H_0\colon \theta^{DM}(u^*) = \theta_0$ is equivalent to testing whether*

$$\exists \delta \in \mathbb{R}^{d_x}\ s.t.\ \widetilde{A}^{DM}_{(\cdot,1)} \left(\theta_0 - u^*_{1,1} P(C=1, Y_1^*=1) + u^*_{0,1} P(C=0)\right) + \widetilde{A}^{DM}_{(\cdot,-1)} \delta \leq \begin{pmatrix} -\underline{P}_{C,Y_1^*}(0,1) \\ \overline{P}_{C,Y_1^*}(0,1) \end{pmatrix},$$

where $\underline{P}_{C,Y_1^*}(0,1), \overline{P}_{C,Y_1^*}(0,1)$ are the d_x-dimensional vectors of lower and upper bounds on $P_{C,Y^*}(C=0, Y_1^*=1 \mid X=x)$ respectively, and $\widetilde{A}^{DM}$ is a known matrix.

C.6 The policymaker's first-best decision rule

For a binary outcome $Y^* = Y_1^* \in \{0,1\}$, consider a policymaker with a linear social welfare function $u^*_{1,1} y_1^* c + u^*_{0,1}(1-y_1^*)(1-c)$ with $u^*_{0,1} < 0, u^*_{1,1} < 0$ as in Section 6 of the main text. I construct an algorithmic decision rule based on analyzing how the policymaker would make choices herself in the binary screening decision. Rambachan et al. (2021) refer to this as the "first-best problem" in their analysis of algorithmic decision rules.

Due to the missing data problem, the conditional probability of $Y_1^*=1$ given the characteristics is partially identified and I assume the policymaker adopts a max-min evaluation

criterion to evaluate decision rules. Let $\tilde{\pi}(x) \in [0,1]$ denote the probability the policymaker selects $C = 1$ given $X = x$. At each $x \in \mathcal{X}$, the policymaker then chooses $\tilde{\pi}(x)$ to maximize

$$\min_{\widetilde{P}(1|x)} \tilde{\pi}(x)\widetilde{P}(1 \mid x)u^*_{1,1} + (1 - \tilde{\pi}(x))(1 - \widetilde{P}(1 \mid x))u^*_{0,1}$$
$$\text{s.t. } \underline{P}(1 \mid x) \leq \widetilde{P}(1 \mid x) \leq \overline{P}(1 \mid x).$$

Proposition C.8. *Assume a binary outcome $Y^* = Y^*_1 \in \{0,1\}$. Consider a policymaker with linear social welfare function $u^*_{0,1} < 0$, $u^*_{1,1} < 0$, who chooses $\tilde{\pi}(x) \in [0,1]$ to maximize worst-case expected social welfare. Defining $\tau^*(u^*) := \frac{u^*_{0,1}}{u^*_{0,1}+u^*_{1,1}} = |u^*_{0,1}|$, her max-min decision rule is*

$$\tilde{\pi}(x) = \begin{cases} 1 \text{ if } \overline{P}(1 \mid x) \leq \tau^*, \\ 0 \text{ if } \underline{P}(1 \mid x) \geq \tau^*, \\ \tau^* \text{ if } \underline{P}(1 \mid x) < \tau^* < \overline{P}(1 \mid x). \end{cases}$$

The policymaker makes choices based on a threshold rule, where the threshold τ^* depends on the relative costs of ex-post errors under the social welfare function. If the upper bound on the probability of $Y^*_1 = 1$ conditional on the characteristics is sufficiently low, then the policymaker chooses $C = 1$ with probability one. If the lower bound on the probability of $Y^* = 1$ is sufficiently high, then the policymaker chooses $C = 0$ with probability one. Otherwise, if the identified set for $P(Y^*_1 = 1 \mid X = x)$ contains the threshold τ^*, the policymaker randomizes and selects $C = 1$ with probability exactly equal to τ^*.

In my empirical analysis in Section 6, I evaluate the choices of judges against this first-best decision rule applied to each cell of included characteristics X_I and each decile of predicted risk $D(X)$. The bounds on the probability defendants would fail to appear in court ($Y^*_1 = 1$) conditional on the characteristics is constructed using the quasi-random assignment of judges as discussed in Section 5.3, and the threshold τ^* varies as the social welfare function $u^*_{0,1}, u^*_{1,1}$ varies. I construct the decision rule using only data from the held-out judges, and treat it as fixed.

Finally, for any particular choice of the social welfare function $u^*_{0,1}, u^*_{1,1}$, I verify whether the resulting algorithmic decision rule is consistent with expected utility maximization at accurate beliefs without private information (since the algorithmic decision rule is only based on observable characteristics). This follows in the spirit of Caplin, Martin and Marx (2022)'s analysis of a pneumonia diagnosis algorithm. Following the logic of the proof of Theorem 3.1, verifying whether the algorithmic decision rule is consistent with expected utility maximization at accurate beliefs is equivalent to checking whether there exists some true conditional outcome probability $P(1 \mid x)$ satisfying the researcher's bounds such that

$$P(1 \mid x) \leq \tau^* \text{ for all } x \text{ s.t. } \tilde{\pi}(x) > 0$$
$$\tau^* \leq P(1 \mid x) \text{ for all } x \text{ s.t. } \tilde{\pi}(x) < 1,$$

for the implied threshold $\tau^* = \frac{u^*_{0,1}}{u^*_{0,1}+u^*_{1,1}} = |u^*_{0,1}|$. For the first-best decision rule defined over each payoff each cell of included characteristics X_I and each decile of predicted risk $D(X)$ and the bounds constructed on the true outcome probability based on the quasi-random

assignment of judges, these inequalities are satisfied for each judge and each social welfare function $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$ and $u^*_{1,1} = -1/|1+\tilde{u}|$ I consider in Section 6. The algorithmic decision rules are, therefore, consistent with expected utility maximization at accurate beliefs at the corresponding social welfare function.

C.7 Proofs of additional results for the econometric framework

C.7.1 Proof of Proposition C.4

Proof. The researcher's bounds on the unobserved conditional outcome probabilities implies bounds on $\widetilde{P}(1 \mid x) \in \mathcal{H}(P(1 \mid x); \mathcal{B}_x)$ as discussed in Section 2.2 of the main text. The result then immediately follows from (23), taking the maximum and minimum over $P(1 \mid x)$ that are consistent with the researcher's bounds. □

C.7.2 Proof of Proposition C.5

Proof. First, rewrite $\theta(\tilde{\pi}, u^*)$ as

$$\beta(\tilde{\pi}, u^*) + \ell^\intercal(\tilde{\pi}, u^*) P_{C,Y_1^*}(1,1) + \ell^\intercal(\tilde{\pi}, u^*) P_{C,Y_1^*}(0,1),$$

where $\ell^\intercal(\tilde{\pi}, u^*)$ is defined in the statement of the proposition and $P_{C,Y_1^*}(1,1)$, $P_{C,Y_1^*}(0,1)$ are the d_x vectors whose elements are $P(C=1, Y_1^* \mid X = x), P(C=0, Y_1^* = 1 \mid X = x)$ respectively. The null hypothesis $H_0 : \theta(\tilde{\pi}, u^*) = \theta_0$ is equivalent to the null hypothesis that there exists $\widetilde{P}_{C,Y_1^*}(0,1)$ satisfying

$$\ell^\intercal(\tilde{\pi}, u^*) \widetilde{P}_{C,Y_1^*}(0,1) = \theta(\tilde{\pi}, u^*) - \beta(\tilde{\pi}, u^*) - \ell^\intercal(\tilde{\pi}, u^*) P_{C,Y_1^*}(1,1)$$
$$\underline{P}(C=0, Y_1^*=1 \mid X=x) \leq \widetilde{P}(C=0, Y_1^*=1 \mid X=x) \leq \overline{P}(C=0, Y_1^*=1 \mid X=x) \text{ for all } x \in \mathcal{X}.$$

We can express the bounds as $A\widetilde{P}_{C,Y_1^*}(0,1) \leq \begin{pmatrix} -\underline{P}_{C,Y_1^*}(0,1) \\ \overline{P}_{C,Y_1^*}(0,1) \end{pmatrix}$, where $A = \begin{pmatrix} -I \\ I \end{pmatrix}$ is a known matrix and $\underline{P}_{C,Y_1^*}(0,1), \overline{P}_{C,Y_1^*}(0,1)$ are the d_x vectors of lower and upper bounds respectively. Therefore, the null hypothesis $H_0 : \theta(\tilde{\pi}, u^*) = \theta_0$ is equivalent to the null hypothesis

$$\exists \widetilde{P}_{C,Y_1^*}(0,1) \text{ satisfying } \ell^\intercal(\tilde{\pi}, u^*) \widetilde{P}_{C,Y_1^*}(0,1) = \theta_0 - \beta(\tilde{\pi}, u^*) - \ell^\intercal(\tilde{\pi}, u^*) P_{C,Y_1^*}(1,1) \text{ and}$$
$$A\widetilde{P}_{C,Y_1^*}(0,1) \leq \begin{pmatrix} -\underline{P}_{C,Y_1^*}(0,1) \\ \overline{P}_{C,Y_1^*}(0,1) \end{pmatrix}.$$

Next, we apply a change of basis argument. Define the full rank matrix Γ, whose first row is equal to $\ell^\intercal(\tilde{\pi}, u^*)$. The null hypothesis $H_0 : \theta(\tilde{\pi}, u^*) = \theta_0$ can be further rewritten as

$$\exists \widetilde{P}_{C,Y_1^*}(0,1) \text{ satisfying } A\Gamma^{-1}\Gamma\widetilde{P}_{C,Y_1^*}(0,1) \leq \begin{pmatrix} -\underline{P}_{C,Y_1^*}(0,1) \\ \overline{P}_{C,Y_1^*}(0,1) \end{pmatrix},$$

where $\Gamma\widetilde{P}_{C,Y_1^*}(0,1) = \begin{pmatrix} \Gamma_{(1,\cdot)}\widetilde{P}_{C,Y_1^*}(0,1) \\ \Gamma_{(-1,\cdot)}\widetilde{P}_{C,Y_1^*}(0,1) \end{pmatrix} = \begin{pmatrix} \theta_0 - \beta(\tilde{\pi}, u^*) - \ell^\intercal(\tilde{\pi}, u^*) P_{C,Y_1^*}(1,1) \\ \delta \end{pmatrix}$ defining

$\delta = \Gamma_{(-1,\cdot)}\widetilde{P}_{C,Y_1^*}(0,1)$ and $\widetilde{A} = A\Gamma^{-1}$. The result then follows immediately with some algebra. □

C.7.3 Proof of Proposition C.6

Proof. The proof follows the same argument as the proof of Proposition C.4. □

C.7.4 Proof of Proposition C.7

Proof. As notation, let $\widetilde{P}_{C,Y_1^*}(0,1 \mid x) := \widetilde{P}(C=0, Y_1^*=1 \mid X=x)$ and let $\widetilde{P}_{C,Y_1^*}(0,1)$ denote the d_x-dimensional vector with entries equal to $\widetilde{P}(C=0, Y_1^* \mid X=x)$. From the definition of $\theta^{DM}(u^*)$, the null hypothesis $H_0 : \theta^{DM}(u^*) = \theta_0$ is equivalent to the null hypothesis that there exists $\widetilde{P}_{C,Y_1^*}(0,1)$ satisfying

$$-u_{0,1}^* \sum_{x\in\mathcal{X}} \widetilde{P}_{C,Y_1^*}(0,1 \mid x)P(X=x) = \theta_0 - u_{1,1}^* P(C=1, Y_1^*=1) - u_{0,1}^* P(C=0)$$
$$\underline{P}(C=0, Y_1^*=1 \mid X=x) \leq \widetilde{P}_{C,Y^*}(0,1 \mid x) \leq \overline{P}(C=0, Y_1^*=1 \mid X=x) \text{ for all } x \in \mathcal{X}.$$

We can express these bounds in the form $A\widetilde{P}_{C,Y_1^*}(0,1) \leq \begin{pmatrix} -\underline{P}_{C,Y_1^*}(0,1) \\ \overline{P}_{C,Y_1^*}(0,1) \end{pmatrix}$, $A = \begin{pmatrix} -I \\ I \end{pmatrix}$ is a known matrix. Therefore, defining $\ell(u^*)$ to be the d_x dimensional vector with entries $-u_{0,1}^* P(X=x)$, the null hypothesis $H_0 : \theta^{DM}(u^*) = \theta_0$ is therefore equivalent to the null hypothesis

$$\exists \widetilde{P}_{C,Y_1^*}(0,1) \text{ satisfying } \ell^\intercal(u^*)\widetilde{P}_{C,Y_1^*}(0,1) = \theta_0 - u_{1,1}^* P(C=1, Y_1^*=1) - u_{0,1}^* P(C=0) \text{ and}$$
$$A\widetilde{P}_{C,Y_1^*}(0,1) \leq \begin{pmatrix} -\underline{P}_{C,Y_1^*}(0,1) \\ \overline{P}_{C,Y_1^*}(0,1) \end{pmatrix}.$$

Next, we apply a change of basis argument. Define the full rank matrix Γ, whose first row is equal to $\ell^\intercal(u^*)$. The null hypothesis $H_0 : \theta^{DM}(u^*) = \theta_0$ can be further rewritten as

$$\exists \widetilde{P}_{C,Y_1^*}(0,1) \text{ satisfying } A\Gamma^{-1}\Gamma\widetilde{P}(0,1) \leq \begin{pmatrix} -\underline{P}_{C,Y_1^*}(0,1) \\ \overline{P}_{C,Y_1^*}(0,1) \end{pmatrix},$$

where $\Gamma\widetilde{P}_{C,Y_1^*}(0,1) = \begin{pmatrix} \Gamma_{(1,\cdot)}\widetilde{P}_{C,Y_1^*}(0,1) \\ \Gamma_{(-1,\cdot)}\widetilde{P}_{C,Y_1^*}(0,1) \end{pmatrix} = \begin{pmatrix} \theta_0 - u_{1,1}^* P(C=1, Y_1^*=1) - u_{0,1}^* P(C=0) \\ \delta \end{pmatrix}$ defining $\delta = \Gamma_{(-1,\cdot)}\widetilde{P}_{C,Y_1^*}(0,1)$ and $\widetilde{A} = A\Gamma^{-1}$. The result then follows immediately with some algebra. □

C.7.5 Proof of Proposition C.8

Proof. To show this result, I consider cases for each $x \in \mathcal{X}$.

Case 1: Suppose $\overline{P}(Y_1^*=1 \mid X=x) \leq \tau^*$. In this case,

$$P(Y_1^*=1 \mid X=x)u_{1,1}^* \geq P(Y^*=0 \mid X=x)u_{0,1}^*$$

for all $P(Y_1^* = 1 \mid X = x)$ satisfying $\underline{P}(Y_1^* = 1 \mid X = x) \leq P(Y^* = 1 \mid X = x) \leq \overline{P}(Y_1^* = 1 \mid X = x)$. Therefore, it is optimal to set $\tilde{\pi}(x) = 1$.

Case 2: Suppose $\underline{P}(Y_1^* = 1 \mid X = x) \geq \tau^*$. In this case,

$$P(Y_1^* = 1 \mid X = x)u_{1,1}^* \leq P(Y_1^* = 0 \mid X = x)u_{0,1}^*$$

for all $P(Y_1^* = 1 \mid X = x)$ satisfying $\underline{P}(Y_1^* = 1 \mid X = x) \leq P(Y_1^* = 1 \mid X = x) \leq \overline{P}(Y_1^* = 1 \mid X = x)$. Therefore, it is optimal to set $\tilde{\pi}(x) = 0$.

Case 3: Suppose $\underline{P}(Y_1^* = 1 \mid X = x) < \tau^* < \overline{P}(Y_1^* = 1 \mid X = x)$. First, notice $\tilde{\pi}(x) = \tau^*$ delivers constant expected payoffs for all $P(Y_1^* = 1 \mid X = x)$ satisfying $\underline{P}(Y_1^* = 1 \mid X = x) \leq P(Y_1^* = 1 \mid X = x) \leq \overline{P}(Y_1^* = 1 \mid X = x)$. As a function of $P(Y_1^* = 1 \mid X = x)$ and $\tilde{\pi}(x)$, expected social welfare equals

$$\tilde{\pi}(x)P(Y_1^* = 1 \mid X = x)u_{1,1}^* + (1 - \tilde{\pi}(x))P(Y_1^* = 0 \mid X = x)u_{0,1}^*.$$

The derivative with respect to $P(Y_1^* = 1 \mid X = x)$ equals $\tilde{\pi}(x)u_{1,1}^* - (1 - \tilde{\pi}(x))u_{0,1}^*$, which equals zero if $\tilde{\pi}(x) = \tau^*$. Moreover, worst case expected social welfare at $\tilde{\pi}(x) = \tau^*$ is equal to the constant $\frac{u_{0,1}^* u_{1,1}^*}{u_{0,1}^* + u_{1,1}^*}$. I show that any other choice of $\tilde{\pi}(x)$ delivers strictly lower worst-case expected social welfare in this case.

Consider any $\tilde{\pi}(x) < \tau^*$. At this choice, expected social welfare is minimized at $\underline{P}(Y_1^* = 1 \mid X = x)$. But, at $\underline{P}(Y_1^* = 1 \mid X = x)$, the derivative of expected social welfare with respect to $\tilde{\pi}(x)$ equals $\underline{P}(Y_1^* = 1 \mid X = x)u_{1,1}^* - (1 - \underline{P}(Y_1^* = 1 \mid X = x))u_{0,1}^*$, which is strictly positive since $\underline{P}(Y_1^* = 1 \mid X = x) < \tau^*$. This implies that

$$\tilde{\pi}(x)\underline{P}(Y_1^* = 1 \mid X = x)u_{1,1}^* + (1 - \tilde{\pi}(x))(1 - \underline{P}(Y_1^* = 1 \mid X = x))u_{0,1}^* <$$

$$\tau^* \underline{P}(Y_1^* = 1 \mid X = x)u_{1,1}^* + (1 - \tau^*)(1 - \underline{P}(Y^* = 1 \mid X = x))u_{0,1}^* = \frac{u_{0,1}^* u_{1,1}^*}{u_{0,1}^* + u_{1,1}^*}.$$

Therefore, worst-case expected social welfare for any $\tilde{\pi}(x) < \tau^*$ is strictly less than worst-case expected social welfare at $\tilde{\pi}(x) = \tau^*$.

Consider any $\tilde{\pi}(x) > \tau^*$. At this choice, expected social welfare is minimized at $\overline{P}(Y_1^* = 1 \mid X = x)$. But, at $\overline{P}(Y_1^* = 1 \mid X = x)$, the derivative of expected social welfare with respect to $\tilde{\pi}(w, x)$ equals $\overline{P}(Y_1^* = 1 \mid X = x)u_{1,1}^* - (1 - \overline{P}(Y^* = 1 \mid X = x))u_{0,1}^*$, which is strictly negative since $\overline{P}(Y_1^* = 1 \mid X = x) > \tau^*$. This implies that

$$\tilde{\pi}(x)\overline{P}(Y_1^* = 1 \mid X = x)u_{1,1}^* + (1 - \tilde{\pi}(x))(1 - \overline{P}(Y_1^* = 1 \mid X = x))u_{0,1}^* <$$

$$\tau^* \overline{P}(Y_1^* = 1 \mid X = x)u_{1,1}^* + (1 - \tau^*)(1 - \overline{P}(Y_1^* = 1 \mid X = x))u_{0,1}^* = \frac{u_{0,1}^* u_{1,1}^*}{u_{0,1}^* + u_{1,1}^*}.$$

Therefore, worst-case expected social welfare for any $\tilde{\pi}(x) > \tau^*$ is strictly less than worst-case expected social welfare at $\tilde{\pi}(x) = \tau^*$. □

D Additional results for expected utility maximization after dimension reduction

In this section, I show how the characterization results for the magnitudes of systematic prediction mistakes and ways in which the decision maker's beliefs are systematically biased can suitably modified to account for the dimension reduction discussed in Section 5.2.

D.1 Approximate expected utility maximization after dimension reduction

As in the main text, let $D\colon \mathcal{X} \to \{1, \dots, N_d\}$ partition the observable characteristics X into level sets $\{x \in \mathcal{X}\colon D(x) = d\}$. For a treatment assignment problem, Theorem B.2 implies that if the decision maker's choices are consistent with approximate expected utility maximization, then their choices satisfy a system of implied revealed preference inequalities over the coarsening $D(\cdot)$. This follows from Lemma B.2 and the same iterated expectations argument as in the proof of Proposition B.2. I omit the proof for brevity.

Proposition D.1. *Assume $0 < \pi_j(x) < 1$ for all $c_j \in \{c_1, \dots, c_J\}$ and $x \in \mathcal{X}$. Suppose the decision maker's choices are consistent with approximate expected utility maximization at $u \in \mathcal{U}$ and $\epsilon = \{\epsilon(x) \geq 0\colon x \in \mathcal{X}\}$. Then, for each $x_I \in \mathcal{X}_I$, $d \in \{1, \dots, N_d\}$, $c_j \in \{c_1, \dots, c_J\}$ and $c' \neq c_j$,*

$$\sum_{\vec{y} \in \mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid (x_I, d)) u(c_j, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid (x_I, d)) u(c', \vec{y}; x_I) - \bar{\epsilon}(x_I, d),$$

where

$$\widetilde{P}(c_j, \vec{y} \mid (x_I, d)) = \sum_{x_E\colon D(x_I, x_E) = d} \widetilde{P}(c_j, \vec{y} \mid (x_I, x_E)) \frac{P(X_E = x_E \mid X_I = x_I)}{P(D(X_I, X_E) = d \mid X_I = x_I)},$$

$$\pi_j(x_I, d) = \sum_{x_E\colon D(x_I, x_E) = d} \pi_j(x_I, x_E) \frac{P(X_E = x_E \mid X_I = x_I)}{P(D(X_I, X_E) = d \mid X_I = x_I)},$$

$$\widetilde{P}_j(\vec{y} \mid (x_I, d)) = \widetilde{P}(c_j, \vec{y} \mid (x_I, d)) / \pi_j(x_I, d),$$

$$\bar{\epsilon}(x_I, d) = \pi_j(x_I, d)^{-1} \sum_{x_E\colon D(x_I, x_E) = d} \epsilon(x_I, x_E) \pi_j(x_I, x_E) \frac{P(X_E = x_E \mid X_I = x_I)}{P(D(X_I, X_E) = d \mid X_I = x_I)}.$$

Corollary D.1. *Suppose $0 < \pi_1(x) < 1$ for all $x \in \mathcal{X}$. If the decision maker's choices approximately maximize expected utility at some linear utility function and $\epsilon = \{\epsilon(x) \geq 0\colon x \in \mathcal{X}\}$, then there exists $\bar{\epsilon}(x_I, d) \geq 0$ such that, for all pairs (x_I, d), (x_I, d'),*

$$\mu_1(x, d) - \overline{\mu}_0(x, d') - \bar{\epsilon}(x, d) - \bar{\epsilon}(x, d') \leq 0.$$

I next show how the coarsening affects the interpretation of the lower bound on the expected utility cost of systematic prediction mistakes. Define the worst-case cost of systematic prediction mistakes over $D(\cdot)$ as $\underline{\mathcal{E}}^*(D) = \sum_{x_I, d} P(x_I, d) \epsilon^*(x_I, d)$, where $\epsilon^*(x)$ is an optimal solution to the linear program defined in (4) of the main text, and $\epsilon^*(x_I, d) =$

$\max_{x_E:\, D(x_I,x_E)=d} \epsilon^*(x_I, x_E)$. That is, $\underline{\mathcal{E}}^*(D)$ is the worst-case cost of systematic prediction mistakes to the decision maker over the partition $D(\cdot)$ since it applies the largest misranking within each cell of the partition over the entire partition. By construction, $\underline{\mathcal{E}}^*(D) \geq \underline{\mathcal{E}}$. Consider the optimal value of the linear program

$$\begin{aligned}\underline{\mathcal{E}}(D) := \min_{\epsilon(x_I,d)} & \sum_{x_I,d} P(x_I,d)\epsilon(x_I, D(x)) \\ \text{s.t. } & \epsilon(x_I,d) \geq 0 \text{ for all } x \in \mathcal{X}, \\ & \mu_1(x_I,d) - \overline{\mu}_0(x_I,d') - \epsilon(x_I,d) - \epsilon(x_I,d') \leq 0 \text{ for all pairs } (x_I,d),(x_I,d').\end{aligned}$$

It is immediate that $\underline{\mathcal{E}}(D) \leq \underline{\mathcal{E}}^*(D)$ since $\epsilon^*(x_I,d)$ is feasible in the program, and so $\underline{\mathcal{E}}(D)$ is a valid lower bound on the worst-case expected utility cost to the decision maker over $D(\cdot)$. Furthermore, $\underline{\mathcal{E}}(D) = 0$ if $\underline{\mathcal{E}} = 0$ by construction.

Analogously, define $\underline{\lambda}^*(D) = \sum_{x_I,d} P(x_I,d)1\{\epsilon^*(x_I,d) > 0\}$ to be the worst-case share of systematic prediction mistakes over $D(\cdot)$. By construction, $\underline{\lambda} \leq \underline{\lambda}^*(D)$. Consider the optimal value of the program

$$\begin{aligned}\underline{\lambda}(D) := \min_{\epsilon(x_I,d)} & \sum_{x_I,d} P(x_I,d)1\{\epsilon(x_I, D(x)) > 0\} \\ \text{s.t. } & \epsilon(x_I,d) \geq 0 \text{ for all } x \in \mathcal{X}, \\ & \mu_1(x_I,d) - \overline{\mu}_0(x_I,d') - \epsilon(x_I,d) - \epsilon(x_I,d') \leq 0 \text{ for all pairs } (x_I,d),(x_I,d').\end{aligned}$$

Since $\epsilon^*(x_I,d)$ is feasible, it follows that $\underline{\lambda}(D) \leq \underline{\lambda}^*(D)$, and so, $\underline{\lambda}(D)$ is a valid lower bound on the worst-case share of systematic prediction mistakes over $D(\cdot)$.

D.2 Expected utility maximization at inaccurate beliefs after dimension reduction

As in the main text, let $D\colon \mathcal{X} \to \{1,\ldots,N_d\}$ partition the observable characteristics X into level sets $\{x \in \mathcal{X}\colon D(x) = d\}$. For a treatment assignment problem, Theorem B.3 implies that if the decision maker's choices are consistent with expected utility maximization at inaccurate beliefs, then their choices satisfy a system of implied revealed preference inequalities over the coarsening $D(\cdot)$. This follows directly from Lemma B.3 and the same iterated expectations argument as in the proof of Proposition 5.1. I omit the proof for brevity.

Proposition D.2. *Suppose the decision maker's choices are consistent with expected utility maximization behavior at inaccurate beliefs and some utility function $u \in \mathcal{U}$. Then, for each $x_I \in \mathcal{X}_I$, $d \in \{1,\ldots,N_d\}$, $c_j \in \{c_1,\ldots,c_J\}$ and $c' \neq c_j$,*

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x_I, d)u(c, \vec{y}; x_I) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x_I, d)u(c', \vec{y}; x_I),$$

where

$$Q(c,\vec{y}\mid x_I,d)=\left(\sum_{x_E:\ D(x_I,x_E)=d}\widetilde{P}(c\mid\vec{y},(x_I,x_E))Q(\vec{y}\mid(x_I,x_E))P(x_E\mid x_I)\right)/P(D(X_I,X_E)=d\mid X_I=x_I),$$

$$\widetilde{P}(c\mid\vec{y},x)=\frac{\widetilde{P}(\vec{y}\mid c,x)\pi_c(x)}{\sum_{c'\in\mathcal{C}}\widetilde{P}(\vec{y}\mid c',x)\pi_{c'}(x)}.$$

Provided that $P(c,\vec{y}\mid x)>0$ for all $(c,\vec{y})\in\mathcal{C}\times\mathcal{Y}^J$ and $x\in\mathcal{X}$, Proposition D.2 can be recast as checking whether there exists non-negative weights $\omega(c,\vec{y};x_I,d)\geq 0$ satisfying, for all $c_j\in\{c_1,\ldots,c_J\}$ with $c'\neq c_j$ and $x\in\mathcal{X}$,

$$\sum_{\vec{y}\in\mathcal{Y}^J}\omega(c_j,\vec{y};x_I,d)\widetilde{P}(c_j,\vec{y}\mid x_I,D(x)=d)u(c_j,\vec{y};x_I)\geq$$

$$\sum_{\vec{y}\in\mathcal{Y}^J}\omega(c_j,\vec{y};x_I,d)\widetilde{P}(c_j,\vec{y}\mid x_I,D(x)=d)u(c',\vec{y};x_I)$$

and $\mathbb{E}_{\widetilde{P}}\left[\omega(C,\vec{Y};X_I,D(X))\mid X_I=x_I,D(X)=d\right]=1$.

I next apply this result in a screening decision with a binary choice and binary outcome. In this special case, following the same argument as the proof of Proposition 4.1, this result may be applied to derive bounds on the decision maker's reweighed utility threshold through

$$P_1(1\mid x_I,d)\leq\frac{\omega(0,0;x_I,d)u_{0,1}(x_I)}{\omega(0,0;x_I,d)u_{0,1}(x_I)+\omega(1,1;x_I,d)u_{1,1}(x_I)}\leq\overline{P}_0(1\mid x_I,d),\tag{24}$$

where $P_c(y^*\mid x_I,d):=P(Y^*=y^*\mid C=c,X_I=x_I,D(X)=d)$. Next, define $M=1\{C=0,Y^*=0\}+1\{C=1,Y^*=1\}$, $\tau(x_I,d)=\frac{\omega(0,0;x_I,d)u_{0,1}(x_I)}{\omega(0,0;x_I,d)u_{0,1}(x_I)+\omega(1,1;x_I,d)u_{1,1}(x_I)}$. Examining $x_I\in\mathcal{X}_I$, $d,d'\in\{1,\ldots,N_d\}$, we arrive at

$$\frac{(1-\tau(x_I,d))/\tau(x_I,d)}{(1-\tau(x_I,d'))/\tau(x_I,d')}=\frac{\frac{Q(C=1,Y^*=1|M=1,x_I,d)/Q(C=0,Y^*=0|M=1,x_I,d)}{Q(C=1,Y^*=1|M=1,x_I,d')/Q(C=0,Y^*=0|M=1,x_I,d')}}{\frac{P(C=1,Y^*=1|M=1,x_I,d)/P(C=0,Y^*=0|M=1,x_I,d)}{P(C=1,Y^*=1|M=1,x_I,d')/P(C=0,Y^*=0|M=1,x_I,d')}}.\tag{25}$$

By examining values in the identified set of reweighted utility thresholds defined on the coarsened characteristic space, bounds may be constructed on a parameter that summarizes the decision maker's beliefs about their own "ex-post errors." That is, how does the decision maker's belief about the relative probability of choosing $C=0$ and outcome $Y^*=0$ occurring vs. choosing $C=1$ and outcome $Y^*=1$ occurring compare to the true probability? If these bounds lie everywhere below one, then the decision maker's beliefs about their own ex-post errors are underreacting to variation in risk across the cells (x_I,d) and (x_I,d'). If these bounds lie everywhere above one, then the decision maker's beliefs about their own ex-post errors are overreacting.

D.3 Constructing coarsened characteristics in screening decisions via out-of-sample prediction

In the empirical application to pretrial release decisions, I construct the partition $D(\cdot)$ of the observed characteristics using supervised machine learning methods that predict the outcome $\bar{Y}^*$ on the pretrial release decisions of other judges. Given an estimated prediction function $\hat{f}\colon \mathcal{X} \to [0, K]$, define $D(\cdot)$ by binning the characteristics X into percentiles of predicted risk within each $x_I \in \mathcal{X}_I$. Provided the prediction function $\hat{f}(\cdot)$ performs well out-of-sample in the sense that it equals the true conditional expectation $\mu(x) := \mathbb{E}[\bar{Y}^* \mid X = x]$ and the excluded characteristics X_E only enter the decision maker's information set through their initial beliefs, then the inequalities in Proposition 5.1 continue to sharply characterize expected utility maximization at accurate beliefs in a screening decision.

Proposition D.3. *Assume $\hat{f}(x) = \mu(x)$, $D(x)$ is defined as the level sets of $\hat{f}(x)$, and $\mu(X)$ is sufficient for X_E in the decision maker's behavior, meaning $V \mid \{Y^*, X_I, X_E\} \sim V \mid \{Y^*, X_I, \mu(X)\}$ and $C \mid \{V, X_I, X_E\} \sim C \mid \{V, X_I, \mu(X)\}$ under $Q(\cdot)$. Then the decision maker's choices are consistent with expected utility maximization at some linear utility function if and only if, for all $x_I \in \mathcal{X}_I$, Equation (8) is satisfied.*

Proof. Under the restriction $V \mid \{Y^*, X\} \sim V \mid \{Y^*, X_I, \mu(X)\}$ and $C \mid \{V, X\} \sim C \mid \{V, X_I, \mu(X)\}$, the expected utility maximization model $(X, V, C, Y^*) \sim Q$ is equivalent to a joint distribution $(X, V, C, Y^*) \sim \tilde{Q}$ for any utility function $u \in \mathcal{U}$ that factorizes according to $\tilde{Q}(X)\tilde{Q}(Y^* \mid X)\tilde{Q}(V \mid Y^*, \mu(X), X_I)\tilde{Q}(C \mid V, \mu(X), X_I)$. The result then follows by the same argument as the proof of Theorem 3.1. □

Proposition D.3 provides a novel connection between out-of-sample prediction and identification in analyzing systematic prediction mistakes. Provided the excluded characteristics X_E only affect the decision maker's behavior through the true conditional expectation $\mu^*(x)$, the extent to which the inequalities in Equation (8) are non-sharp is driven by how well the estimated prediction function recovers $\mu^*(x)$. Nonetheless, the inequalities in Equation (8) always provide a valid falsification test for accurate beliefs regardless of whether the conditions in Proposition D.3 are satisfied.

E Additional empirical results for the New York City pretrial system

E.1 Defining the outcome to be any pretrial misconduct

Given the stated objectives of the NYC pretrial system, the main text defined the outcome $Y^* = Y_1^* \in \{0, 1\}$ to be whether a defendant would fail to appear in court. As a robustness exercise, I define the outcome of interest $Y^* = Y_1^* \in \{0, 1\}$ to be whether a defendant would commit "any pretrial misconduct" (i.e., either fail to appear in court or be re-arrested for any new crime).

What fraction of judges make prediction mistakes? I test whether the release decisions of each judge in the top 25 are consistent with expected utility maximization behavior at some linear utility function that (i) does not depend on any observable characteristics,

(ii) depends on the defendant's race, (iii) depends on both the defendant's race and age, or (iv) depends on the defendant's race and whether the defendant was charged with a felony offense. Online Appendix Table A1 shows that the pretrial release decisions of at least 64% of judges are inconsistent with expected utility maximization at accurate beliefs about pretrial misconduct risk and some linear utility function satisfying the conjectured exclusion restrictions.

How common and costly are systematic prediction mistakes? I estimate the bound on the share of systematic prediction mistakes about pretrial misconduct risk assuming judges optimize some linear utility function that depends on both the defendant's race and age or the defendant's race and whether the defendant was charged with a felony offense (see Section 4.1.2 for theoretical details). Online Appendix Table A2 summarizes the results across judges whose choices are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level.

I also estimate the bound on the total expected utility cost to judges of their systematic prediction mistakes about pretrial misconduct risk (see Section 4.1.1 for theoretical details) across the same judges. Using the procedure in Online Appendix C.2, I translate these estimated expected utility costs into an equivalent reduction in the fraction of defendants that are released and would commit pretrial misconduct that would produce the same expected utility cost to the judge. For the median judge, this corresponds to an equivalent reduction of 25.06 percentage points when both defendant race and age are allowed to directly affect utility and 25.90 percentage points when defendant race and charge severity are allowed to directly affect utility.

Bounding prediction mistakes based on defendant characteristics Online Appendix Figure A7a reports 95% confidence intervals for the identified set of the implied prediction mistake $\delta(x_I, d)/\delta(x_I, d')$ between the highest d and lowest decile d' of predicted pretrial misconduct risk within each race-by-age X_I cell. Online Appendix Figure A7b plots the 95% confidence intervals for the identified set on the same object within each race-by-felony charge X_I cell. See Section 4.2 for details. Judges appear to systematically underreact to predictable variation in pretrial misconduct risk between defendants at the tails of the pretrial misconduct risk distribution. Whenever these bounds are informative, they lie strictly below one.

Furthermore, among judges whose choices are inconsistent with expected utility maximization behavior at accurate beliefs about pretrial misconduct risk, Online Appendix Table A3 reports the location of the largest studentized misranking and shows the fraction of judges for whom the largest misranking occurs over the tails of the predicted distribution (deciles 1-2, 9-10) or the middle of the predicted risk distribution (deciles 3-8) for black and white defendants respectively. I again find that the largest misrankings mainly occur over defendants that lie in the tails of the predicted risk distribution, and furthermore the majority occur over black defendants at the tails of the predicted risk distribution.

E.2 Identifying prediction mistakes under alternative bounds on the missing data

Section 5 analyzed the pretrial release decisions of judges in New York City by constructing bounds on the failure to appear rate of detained defendants using the quasi-random assignment of judges. I now show how the same analysis may be conducted instead assuming

$$P(Y_1^* = 1 \mid C = 1, X_I = x_I, D(X) = d) \leq P(Y_1^* = 1 \mid C = 0, X_I = x_I, D(X) = d), \quad (26)$$

$$P(Y_1^* = 1 \mid C = 0, X_I = x_I, D(X) = d) \leq (1 + \kappa) P(Y_1^* = 1 \mid C = 0, X_I = x_I, D(X) = d) \quad (27)$$

for some chosen parameter $\kappa \geq 0$. The parameter $\kappa \geq 0$ bounds the failure to appear rate among detained defendants relative to the failure to appear rate among released defendants, and I report results as $\kappa \geq 0$ varies.

Under this bounding assumption, in Section 5.4.3 of the main text, I tested whether the release decisions of each judge in the top 25 are consistent with expected utility maximization behavior at some linear utility function that (i) does not depend on any observable characteristics, (ii) depends on the defendant's race, (iii) depends on both the defendant's race and age, or (iv) depends on the defendant's race and whether the defendant was charged with a felony offense. Figure II reports the fraction of judges in the top 25 for whom we can reject expected utility maximization at accurate beliefs under various assumption on which observable characteristics X_I affect the utility function and varying the choice of $\kappa \geq 0$.

How common and costly are systematic prediction mistakes? I estimate the bound on the share of systematic prediction mistakes about failure to appear risk assuming judges' optimize some linear utility function that depends on both the defendant's race and age or the defendant's race and whether the defendant was charged with a felony offense (see Section 4.1.2). Online Appendix Figure A8 reports the estimated bounds on the share of systematic prediction mistakes across judges whose choices are inconsistent with expected utility maximization at accurate beliefs about failure to appear risk at the nominal 5% level as $\kappa \geq 0$ varies.

I also estimate the bound on the total expected utility costs to judges of their systematic prediction mistakes as $\kappa \geq 0$ varies (see Section 4.1.1). Online Appendix Figure A9a reports the estimated bounds on the total expected utility cost $\underline{\mathcal{E}}$, and Online Appendix Figure A9b translates these estimated total expected utility costs into an equivalent reduction in the fraction of defendants that are released and would fail to appear in court that would produce the same total expected utility cost.

Bounding prediction mistakes based on defendant characteristics I next analyze whether judges over-react or under-react to variation in failure to appear risk based on the observable characteristics. Online Appendix Figure A10a reports 95% confidence intervals for the identified set of values $\delta(x_I, d)/\delta(x_I, d')$ between the highest d and lowest decile d' of predicted risk within each race-by-age cell using the bounds with $\kappa = 2$. Online Appendix Figure A10b plots the same results for each race-by-felony charge cell. See Section 4.2 for theoretical details on the implied prediction mistake $\delta(x_I, d)/\delta(x_I, d')$. As in the main text,

judges appear to underreact to predictable variation in failure to appear risk. Whenever informative, these bounds lie strictly below one.

E.3 The welfare effects of algorithmic decision-making: race-by-felony charge cells

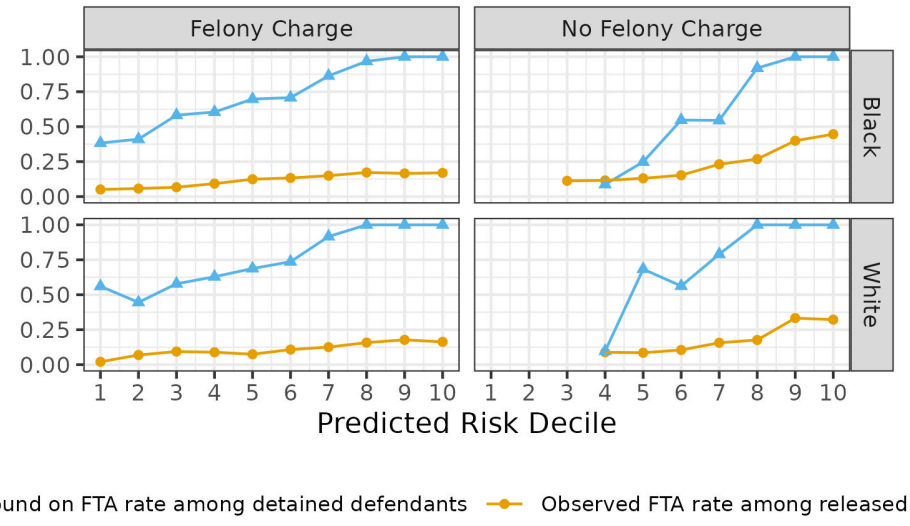
Section 6 compared worst-case expected social welfare under the observed release decisions by judges in New York City against worst-case expected social welfare under counterfactual algorithmic decisions, conducting this exercise over race-by-age cells and deciles of predicted failure to appear risk. I report the results of the same analysis over race-by-felony charge cells and deciles of predicted failure to appear risk for completeness and find analogous results.

Online Appendix Figure A11a plots the improvement in worst-case total expected social welfare under the algorithmic decision rule that fully replaces judges who were found to make systematic prediction mistakes against the release decisions of these judges. For most values of the social welfare function, the algorithmic decision rule dominates the observed choices of these judges, but for social welfare costs of unnecessary detentions ranging over $|\tilde{u}| \in [0.3, 0.7]$ (recall $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|, u^*_{1,1} = -1/|1+\tilde{u}|$), the algorithmic decision rule either leads to no improvement or strictly lowers worst-case expected total social welfare relative to the judges' observed decisions. Online Appendix Figure A11b plots the improvement in worst-case expected social welfare under the algorithmic decision rule that only corrects systematic prediction mistakes at the tails of the predicted failure to appear risk distribution against the observed release decisions of these judges. The algorithmic decision rule that only corrects systematic prediction mistakes weakly dominates the observed release decisions of judges, no matter the value of the social welfare function.

I compare welfare effects of replacing judges whose choices were found to be consistent with expected utility maximization behavior at accurate beliefs about failure to appear risk with algorithmic decision rules. Online Appendix Figure A12 plots the improvement in worst-case total expected social welfare under the algorithmic decision rule that fully replaces these judges against their release decisions. Replacing these judges with algorithmic decision rules may strictly lower worst-case expected social welfare for a range of social welfare costs of unnecessary detentions.

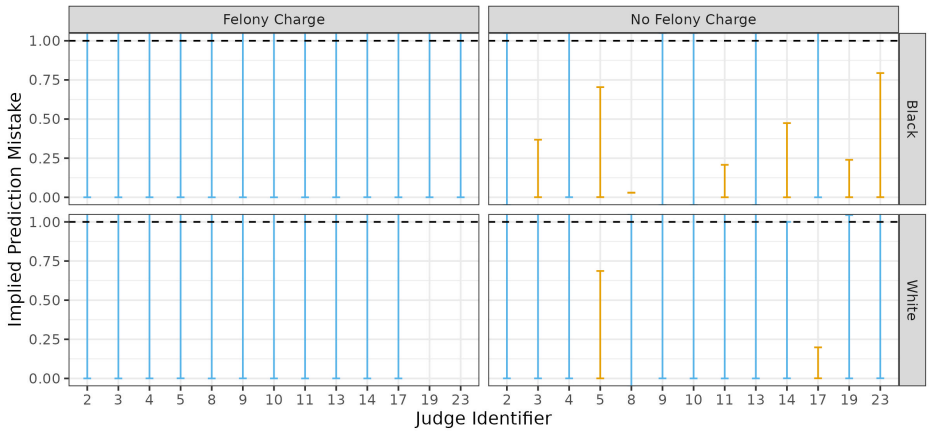
F Online appendix figures

Figure A1: Judge-specific failure to appear rate among released defendants and bound on the failure to appear rate among detained defendants by race-and-felony charge cells.



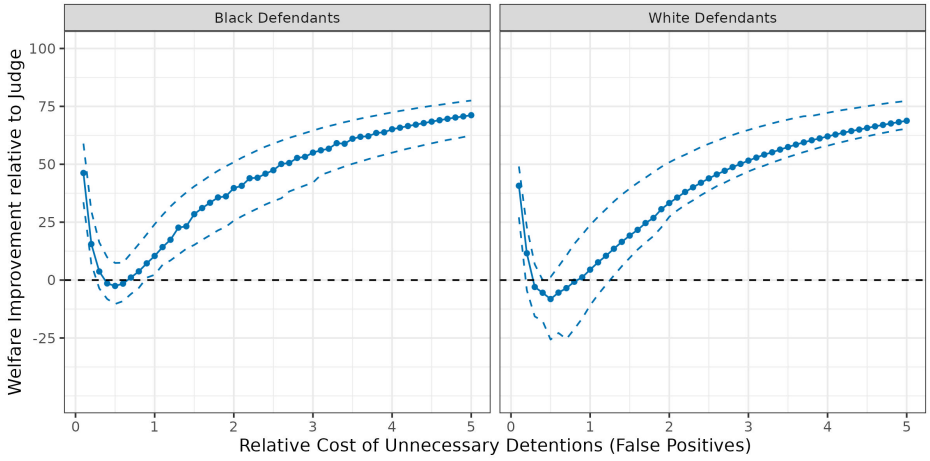
Notes: This figure plots the observed failure to appear rate among released defendants (orange, circles) and the bounds based on the judge leniency for the failure to appear rate among detained defendants (blue, triangles) at each decile of predicted failure to appear risk and race-by-felony charge cell for the judge that heard the most cases in the main estimation sample. The bounds on the failure to appear rate among detained defendants (blue, triangles) are constructed using the most lenient quintile of judges. See Section 5.3 for discussion.

Figure A2: Judge-specific bounds on prediction mistakes between predicted failure to appear risk deciles within each race-by-felony charge cell.



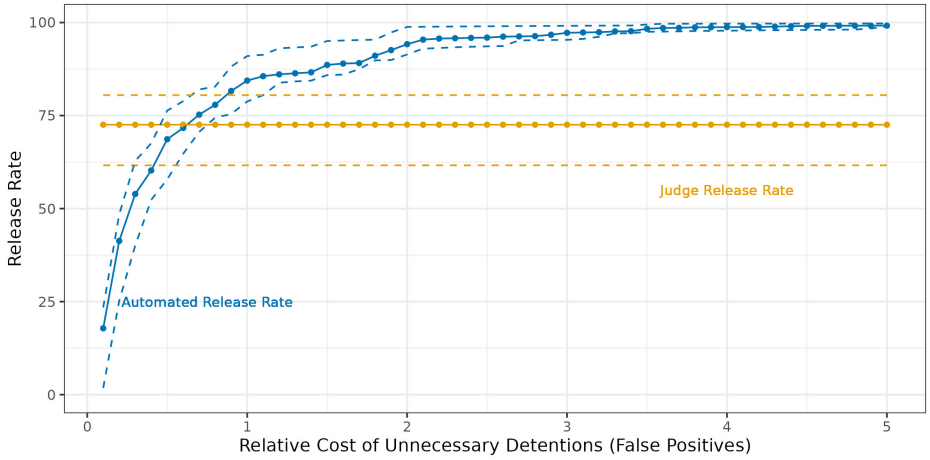
Notes: This figure plots the 95% confidence interval on the implied prediction mistake $\delta(x_I, d)/\delta(x_I, d')$ between the top decile d and bottom decile d' of the predicted failure to appear risk distribution for each judge in the top 25 whose pretrial release decisions are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level (the "unadjusted rejection rate" in the top panel of Table II) and each race-by-felony charge cell. The confidence intervals highlighted in orange show that judges under-react to predictable variation in failure to appear risk from the highest to the lowest decile of predicted failure to appear risk (i.e., the estimated bounds lie below one). See Section 4.2 for theoretical details on the implied prediction mistake and Section 5.6 for the estimation details.

Figure A3: Comparison of algorithmic decision rule against judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs, separately by defendant race.



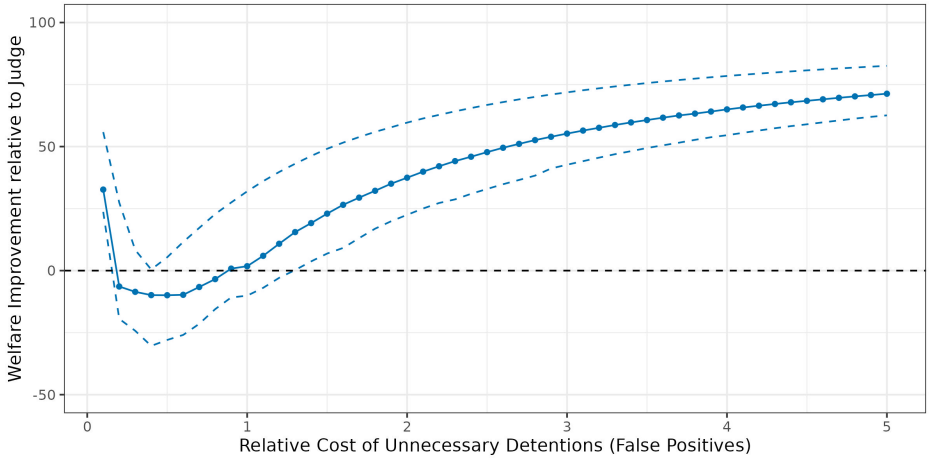
Notes: This figure reports the change in worst-case expected social welfare under the algorithmic decision rule against judges whose pretrial release decisions are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level (the "unadjusted rejection rate" in the top panel of Table II), separately by defendant race. The x-axis plots the relative social welfare cost of detaining a defendant that would not fail to appear in court $|\tilde{u}|$ (i.e., an unnecessary detention), where $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$, $u^*_{1,1} = -1/|1+\tilde{u}|$. The solid line plots the median change across judges that make mistakes, and the dashed lines report the minimum and maximum change across judges. See Section 6 for discussion.

Figure A4: Release rates under algorithmic decision rule relative to the release rates of judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs.



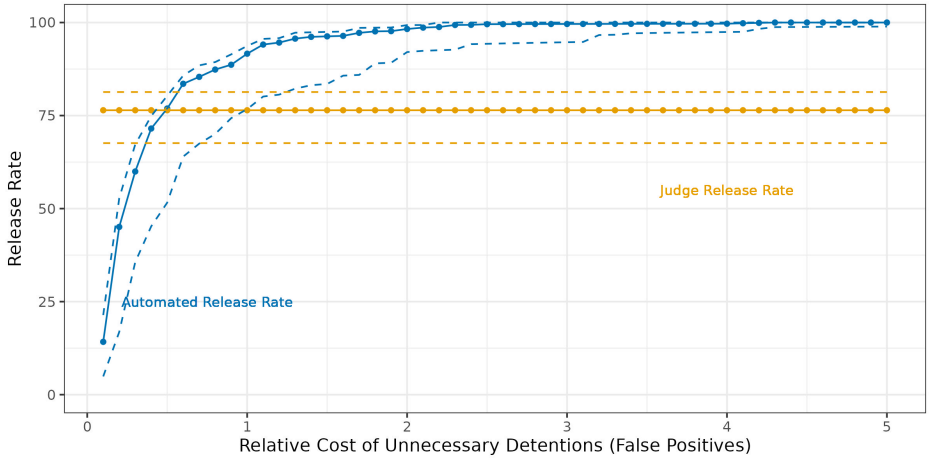
Notes: This figure reports the overall release rate of the algorithmic decision rules against the release rates of judges whose pretrial release decisions are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level (the "unadjusted rejection rate" in the top panel of Table II). These decisions rules are constructed and evaluated over race-by-age cells and deciles of predicted risk. The x-axis plots the relative social welfare cost of detaining a defendant that would not fail to appear in court $|\tilde{u}|$ (i.e., an unnecessary detention), where $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$, $u^*_{1,1} = -1/|1+\tilde{u}|$. The solid line plots the median release rate across judges that make systematic prediction mistakes, and the dashed lines report the minimum and maximum release rates across judges. See Section 6 for discussion.

Figure A5: Comparison of algorithmic decision rule against judges whose choices are consistent with expected utility maximization at accurate beliefs.



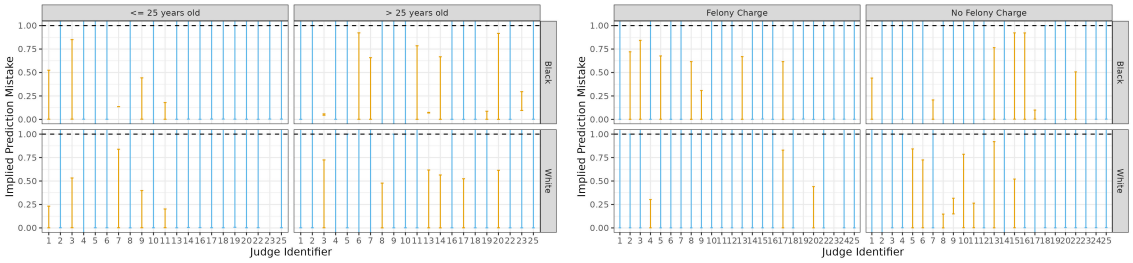
Notes: This figure reports the change in worst-case expected social welfare under the algorithmic decision rule against judges whose pretrial release decisions are consistent with expected utility maximization at accurate beliefs about failure to appear risk at the nominal 5% level (the "unadjusted rejection rate" in the top panel of Table II). The x-axis plots the relative social welfare cost of detaining a defendant that would not fail to appear in court $|\tilde{u}|$ (i.e., an unnecessary detention), where $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$, $u^*_{1,1} = -1/|1+\tilde{u}|$. The solid line plots the median change across judges, and the dashed lines report the minimum and maximum change across judges. See Section 6 for discussion.

Figure A6: Release rates under algorithmic decision rule relative to the release rates of judges whose choices are consistent with expected utility maximization at accurate beliefs.



Notes: This figure reports the release rate of the algorithmic decision rules against the observed release rates among judges whose choices are consistent with expected utility maximization behavior at accurate beliefs at the nominal 5% level (the "unadjusted rejection rate" in the top panel of Table II). The algorithmic decision rules are constructed and evaluated over race-by-age cells and deciles of predicted failure to appear risk. The x-axis plots the relative social welfare cost of detaining a defendant that would not fail to appear in court $|\tilde{u}|$ (i.e., an unnecessary detention), where $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$, $u^*_{1,1} = -1/|1+\tilde{u}|$. The solid line plots the median release rate across judges that do not make systematic prediction mistakes, and the dashed lines report the minimum and maximum release rates across judges. See Section 6 for discussion.

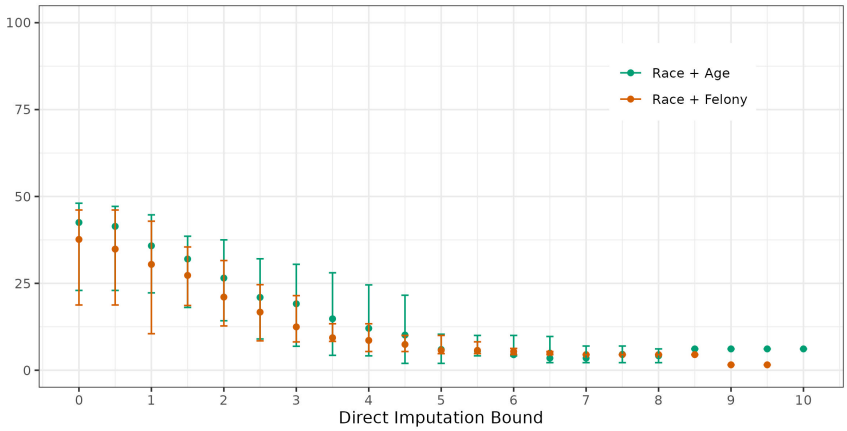
Figure A7: Judge-specific bounds on prediction mistakes about pretrial misconduct risk between predicted risk deciles



(a) Race-by-age X_I cells **(b)** Race-by-felony charge X_I cells

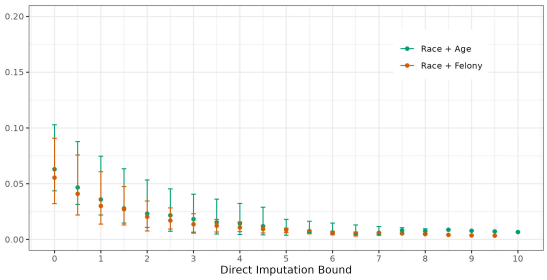
Notes: This figures plots the 95% confidence interval for the identified set on $\delta(x_I, d)/\delta(x_I, d')$ between the highest predicted any pretrial misconduct risk decile d and the lowest predicted any pretrial misconduct risk decile d' within each race-by-age cell and race-by-felony charge cell. I report results for judges' whose choices are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level. The outcome Y_1^* is whether the defendant would commit any pretrial misconduct upon release (i.e., either fail to appear in court or be re-arrested for a new crime). Bounds on the any pretrial misconduct rate among detained defendants are constructed using the judge leniency instrument (see Section 5.3). See Section 4.2 for theoretical details on the implied prediction mistake and Online Appendix E.1 for discussion.

Figure A8: Share of systematic prediction mistakes using alternative bounds on the missing data.

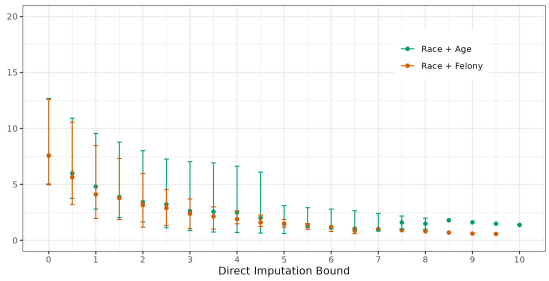


Notes: This figure summarizes the estimated bound on the share of systematic prediction mistakes among judges whose choices are inconsistent with expected utility maximization at accurate beliefs and a linear utility function that depends on both the (i) defendant's race and age, and (ii) defendant's race and whether the defendant was charged with a felony offense. Bounds on the failure to appear rate among detained defendants are constructed using Equation (26) with $\kappa = \{0, 1, \ldots, 10\}$. The lower and upper error bars summarize the minimum and maximum across judges whose choices are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level. The dot summarizes the median estimated bound across those same judges. See Section 4.1.2 for the theoretical details, as well as Section 5.4.3 and Online Appendix E.2 for further discussion.

Figure A9: Total expected utility costs of systematic prediction mistakes using alternative bounds on the missing data.



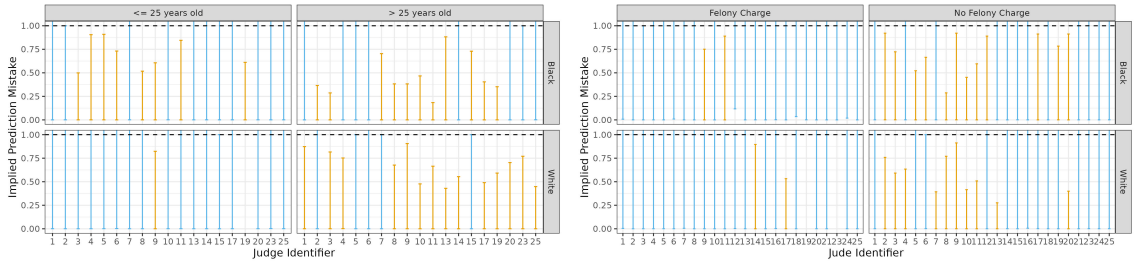
(a) Expected utility units



(b) Equivalent reduction in percentage of ex-post errors $E[Y_1^* \cdot C]$

Notes: This figure summarizes the estimated bounds on the total expected utility costs of systematic prediction mistakes among judges whose choices are inconsistent with expected utility maximization at accurate beliefs and a linear utility function that depends on both the (i) defendant's race and age, and (ii) defendant's race and whether the defendant was charged with a felony offense. Bounds on the failure to appear rate among detained defendants are constructed using Equation (26) with $\kappa = \{0, 1, \ldots, 10\}$. Panel (a) reports the bound on the total expected utility cost $\underline{\mathcal{E}}$, and Panel (b) reports the equivalent reduction in the fraction of defendants that are released and would fail to appear in court that would produce the same total expected utility cost using the procedure described in Appendix C.2. The lower and upper error bars summarize the minimum and maximum across judges whose choices are inconsistent with expected utility maximization at accurate beliefs at the nominal 5% level. The dot summarizes the median estimated bound across those same judges. See Section 4.1.2 for the theoretical details, as well as Section 5.4.3 and Online Appendix E.2 for further discussion.

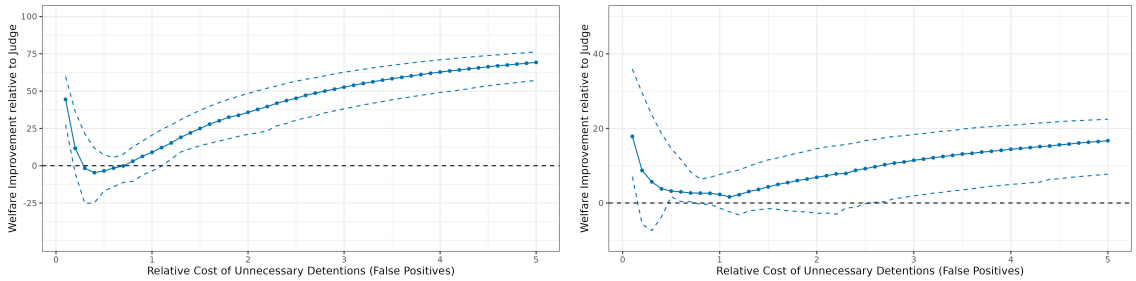
Figure A10: Judge-specific bounds on prediction mistakes between predicted failure to appear risk deciles using alternative bounds on the missing data.



(a) Race-by-age X_I cells **(b)** Race-by-felony charge X_I cells

Notes: This figures plots the 95% confidence interval for the identified set on the implied prediction mistake $\delta(x_I, d)/\delta(x_I, d')$ between the highest predicted failure to appear risk decile d and the lowest predicted failure to appear risk decile d' within each race-by-age cell and race-by-felony charge cell. The bounds on the failure to appear rate among detained defendants are constructed using Equation (26) for $\kappa = 2$ and for each judge in the top 25 whose choices are inconsistent with expected utility maximization behavior at these bounds at the nominal 5% level. See Section 4.2 for theoretical details on the implied prediction mistake, as well as Section 5.4.3 and Online Appendix E.2 for further discussion.

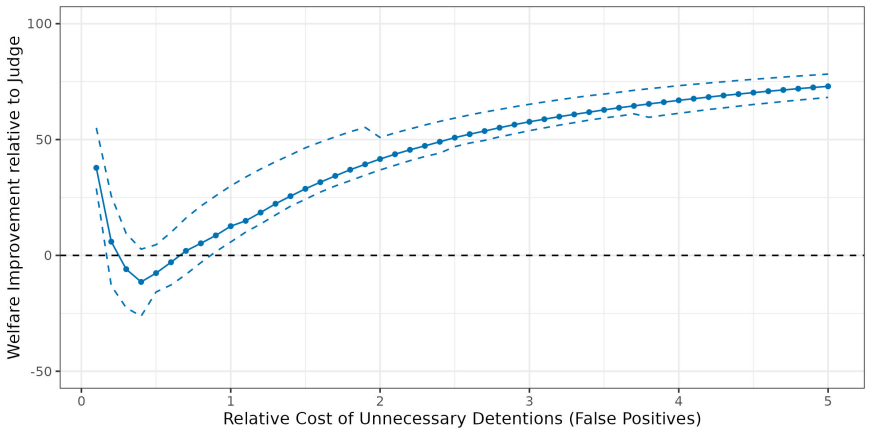
Figure A11: Comparison of algorithmic decision rule against judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs over race-by-felony charge cells.



(a) Algorithmic decision rule that fully replaces judges **(b)** Algorithmic decision rule that corrects prediction mistakes

Notes: This figure reports the change in worst-case total expected social welfare under two algorithmic decision rules against the judge's observed release decisions among judges whose pretrial release decisions are inconsistent with expected utility maximization at accurate beliefs over race-by-felony charge cells. Worst case total expected social welfare under each decision rule is computed by first constructing a 95% confidence interval for total expected social welfare under the decision rule, and reporting smallest value that lies in the confidence interval. These decisions rules are constructed and evaluated over race-by-felony cells and deciles of predicted failure to appear risk. The x-axis plots the relative social welfare cost of detaining a defendant that would not fail to appear in court $|\tilde{u}|$ (i.e., an unnecessary detention), where $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$. The solid line plots the median change across judges that make mistakes, and the dashed lines report the minimum and maximum change across judges. See Section 6 of the main text and Online Appendix E.3 for further details.

Figure A12: Comparison of algorithmic decision rule against judges whose choices are consistent with expected utility maximization at accurate beliefs over race-by-felony charge cells.



Notes: This figure reports the change in worst-case total expected social welfare under the algorithmic decision rule that fully replaces judges against observed release decisions among judges whose choices are consistent with expected utility maximization behavior at accurate beliefs about failure to appear risk. Worst case total expected social welfare under each decision rule is computed by first constructing a 95% confidence interval for total expected social welfare under the decision rule, and reporting smallest value that lies in the confidence interval. These decisions rules are constructed and evaluated over race-by-felony cells and deciles of predicted risk. The x-axis plots the relative social welfare cost of detaining a defendant that would not fail to appear in court $|\tilde{u}|$ (i.e., an unnecessary detention), where $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$. The solid line plots the median change across judges that make mistakes, and the dashed lines report the minimum and maximum change across judges. See Section 6 of the main text and Online Appendix E.3 for further details.

G Online appendix tables

Table A1: Fraction of judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs about pretrial misconduct risk.

	Utility Functions $u(c, y^*; x_I)$			
	No Characteristics	Race	Race + Age	Race + Felony Charge
Adjusted Rejection Rate	76%	72%	64%	92%

Notes: This table summarizes the results for testing for misrankings in the release decisions of each judge in the top 25 at linear utility functions $u(c, y^*; x_I)$ that (i) do not depend on any characteristics, (ii) depend on the defendant's race, (iii) depend on both the defendant's race and age, and (iv) depend on both the defendant's race and whether the defendant was charged with a felony offense. The outcome $Y^* = Y_1^* \in \{0, 1\}$ is whether the defendant would commit any pretrial misconduct (i.e., either fail to appear in court or be re-arrested for a new crime) upon release. Bounds on the any pretrial misconduct rate among detained defendants are constructed using the judge leniency instrument (see Section 5.3). I test the moment inequalities using the conditional least-favorable hybrid test developed in Andrews, Roth and Pakes (2023). I estimate the variance-covariance matrix of the any pretrial misconduct rate among released defendants and upper bounds on the any pretrial misconduct rate among detained defendants using the bootstrap conditional on the included characteristics X_I, predicted risk decile $D(X)$ and leniency quintile instrument Z. The adjusted rejection rate reports the fraction of rejections after a multiple hypothesis testing correction that controls the family-wise error rate at the 5% level. See Online Appendix E.1 for further discussion.

Table A2: Share of systematic prediction mistakes among judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs about pretrial misconduct risk.

	Utility Functions $u(c, y; x_I)$	
	Race and Age	Race and Felony Charge
Unadjusted Rejection Rate	84%	98%
Prediction Mistake Share		
Minimum	54.87%	59.37%
Median	61.19%	71.30%
Maximum	72.55%	80.93%

Notes: This table summarizes the estimated bound on the share of systematic prediction mistakes among judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs about pretrial misconduct risk and utility functions that depend on both the defendant's race and age as well as the defendant's race and whether the defendant was charged with a felony. Among judges' whose choices were inconsistent with expected utility maximization at the 5% level ("unadjusted rejection rate"), I compute the optimal value of the sample analogue to the optimization program (21). See Section 4.1.2 for theoretical details on the bound for the share of systematic prediction mistakes, and Online Appendix E.1 for discussion.

Table A3: Largest misranking among judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs about pretrial misconduct risk.

	Utility Functions $u(c, y; x_I)$	
	Race and Age	Race and Felony Charge
Unadjusted Rejection Rate	84%	98%
White Defendants		
Middle Deciles	0.00%	0.00%
Tail Deciles	4.76%	4.16%
Black Defendants		
Middle Deciles	9.52%	16.66%
Tail Deciles	85.71%	79.16%

Notes: This table summarizes the location of the largest (studentized) misranking in Proposition 5.1 among judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs about pretrial misconduct risk and preferences that depend on both the defendant's race and age as well as the defendant's race and whether the defendant was charged with a felony. The outcome Y_1^* is whether the defendant would commit any pretrial misconduct upon release (i.e., either fail to appear in court or be re-arrested for a new crime). See Online Appendix E.1 for discussion.

Table A4: Summary statistics comparing the main estimation sample and cases heard by the top 25 judges.

	All Defendants		White Defendants		Black Defendants	
	Estimation Sample	Top Judges	Estimation Sample	Top Judges	Estimation Sample	Top Judges
	(1)	(2)	(3)	(4)	(5)	(6)
Released before trial	0.720	0.736	0.757	0.777	0.687	0.699
Defendant Characteristics						
White	0.475	0.481	1.000	1.000	0.000	0.000
Female	0.173	0.173	0.154	0.152	0.190	0.192
Age at Arrest	31.95	31.75	32.03	31.88	31.87	31.63
Arrest Charge						
Number of Charges	1.152	1.167	1.187	1.217	1.119	1.121
Felony Charge	0.372	0.367	0.367	0.356	0.376	0.377
Any Drug Charge	0.253	0.224	0.253	0.217	0.253	0.230
Any DUI Charge	0.047	0.049	0.070	0.072	0.027	0.027
Any Violent Crime Charge	0.375	0.395	0.358	0.379	0.390	0.410
Property Charge	0.130	0.132	0.122	0.123	0.138	0.140
Defendant Priors						
Any FTA	0.516	0.497	0.443	0.419	0.582	0.570
Number of FTAs	2.177	2.034	1.633	1.492	2.670	2.537
Any Misdemeanor Arrest	0.683	0.667	0.615	0.596	0.744	0.734
Any Misdemeanor Conviction	0.383	0.368	0.334	0.315	0.427	0.418
Any Felony Arrest	0.581	0.566	0.503	0.482	0.652	0.644
Any Felony Conviction	0.285	0.271	0.234	0.215	0.331	0.323
Any Violent Felony Arrest	0.398	0.387	0.306	0.292	0.481	0.476
Any Violent Felony Conviction	0.119	0.114	0.084	0.078	0.150	0.147
Total Cases	569,256	243,118	270,704	117,073	298,552	126,045

Notes: This table provides summary statistics for the main estimation sample and the cases heard by the top 25 judges in the New York City pretrial release data for all defendants and separately by defendant race. See Section 5.1 of the main text for further discussion.

Table A5: Summary statistics for released and detained defendants in the main estimation sample and for cases heard by the top 25 judges

	All Defendants		Released Defendants		Detained Defendants	
	Estimation Sample	Top Judges	Estimation Sample	Top Judges	Estimation Sample	Top Judges
	(1)	(2)	(3)	(4)	(5)	(6)
Released before trial	0.720	0.736	1.000	1.000	0.000	0.000
Defendant Characteristics						
White	0.475	0.481	0.499	0.508	0.412	0.407
Female	0.173	0.173	0.199	0.197	0.107	0.106
Age at Arrest	31.95	31.75	31.22	31.20	33.82	33.29
Arrest Charge						
Number of Charges	1.152	1.167	1.148	1.162	1.161	1.182
Felony Charge	0.372	0.367	0.288	0.288	0.588	0.586
Any Drug Charge	0.253	0.224	0.229	0.204	0.314	0.279
Any DUI Charge	0.047	0.049	0.062	0.063	0.010	0.010
Any Violent Crime Charge	0.375	0.395	0.388	0.409	0.341	0.355
Property Charge	0.130	0.132	0.115	0.114	0.171	0.181
Defendant Priors						
Any FTA	0.516	0.497	0.409	0.395	0.793	0.784
Number of FTAs	2.177	2.034	1.362	1.295	4.284	4.103
Any Misdemeanor Arrest	0.683	0.667	0.610	0.598	0.871	0.863
Any Misdemeanor Conviction	0.383	0.368	0.284	0.278	0.637	0.621
Any Felony Arrest	0.581	0.566	0.487	0.477	0.824	0.814
Any Felony Conviction	0.285	0.271	0.200	0.194	0.505	0.487
Any Violent Felony Arrest	0.398	0.387	0.315	0.309	0.614	0.608
Any Violent Felony Conviction	0.119	0.114	0.081	0.080	0.216	0.210
Total Cases	569,256	243,118	410,394	179,143	158,862	63,975

Notes: This table provides summary statistics for the main estimation sample and the cases heard by the top 25 judges in the New York City pretrial release data for all defendants and separately by whether the defendant was released or detained. See Section 5.1 of the main text for discussion.

Table A6: Balance check estimates for the quasi-random assignment of judges by defendant race and age.

	White Defendants		Black Defendants	
	Young (1)	Older (2)	Young (3)	Older (4)
Defendant Characteristics				
Female	−0.00008	0.00017	−0.00007	−0.00005
	(0.00025)	(0.00019)	(0.00024)	(0.00024)
Age	−0.000004	−0.00001	−0.00006	−0.00001
	(0.00004)	(0.00001)	(0.00003)	(0.00001)
Arrest Charge				
Number of Charges	−0.00002	−0.000003	−0.00002	0.00001
	(0.00003)	(0.000005)	(0.00006)	(0.00003)
Felony Charge	0.00002	−0.00024	0.00019	0.00033
	(0.00023)	(0.00019)	(0.00023)	(0.00022)
Any Drug Charge	−0.00033	0.00004	−0.00046	0.00004
	(0.00033)	(0.00022)	(0.00025)	(0.00020)
Any Violent Crime Charge	−0.00025	−0.00010	−0.00016	0.00018
	(0.00026)	(0.00019)	(0.00024)	(0.00018)
Any Property Charge	−0.00005	−0.00046	−0.00017	−0.00045
	(0.00034)	(0.00023)	(0.00031)	(0.00029)
Any DUI Charge	0.00021	0.00042	−0.00160	0.00062
	(0.00045)	(0.00030)	(0.00072)	(0.00044)
Defendant Priors				
Prior FTA	−0.00013	−0.00015	0.00034	−0.00021
	(0.00026)	(0.00021)	(0.00022)	(0.00020)
Prior Misdemeanor Arrest	0.00026	−0.00018	−0.00008	0.00034
	(0.00021)	(0.00017)	(0.00022)	(0.00022)
Prior Felony Arrest	−0.00008	0.00018	0.00035	−0.00025
	(0.00026)	(0.00027)	(0.00030)	(0.00024)
Prior Violent Felony Arrest	−0.00024	−0.00001	−0.00020	−0.00019
	(0.00030)	(0.00023)	(0.00025)	(0.00021)
Prior Misdemeanor Conviction	0.00040	0.00023	0.00040	0.00004
	(0.00029)	(0.00025)	(0.00028)	(0.00018)
Prior Felony Conviction	0.00052	0.00005	−0.00094	−0.00016
	(0.00049)	(0.00019)	(0.00033)	(0.00017)
Prior Violent Felony Conviction	−0.00029	−0.00020	0.00113**	−0.00012
	(0.00077)	(0.00022)	(0.00054)	(0.00021)
Joint p-value	0.85104	0.44370	0.038862	0.16062
Court × Time FE	✓	✓	✓	✓
Cases	99,536	171,168	119,156	179,396

Notes: This table reports OLS estimates for regressions of the constructed judge leniency measure on various defendant and case characteristics in the main estimation sample. These regressions are estimated separately over subsamples defined on the race and age of the defendant, where "young" is defined as less than or equal to 25 years and "old" is defined as older than 25 years. Standard errors, reported in parentheses, are clustered at the defendant and judge level. The joint p-value is based on the F-statistic for whether all defendant and case characteristics are jointly significant. See Section 5.3 of the main text for discussion.

Table A7: Balance check estimates for the quasi-random assignment of judges by defendant race and felony charge.

	White Defendants		Black Defendants	
	Felony Charge (1)	No Felony Charge (2)	Felony Charge (3)	No Felony Charge (4)
Defendant Characteristics				
Female	0.00003	0.00001	−0.00003	−0.00004
	(0.00023)	(0.00021)	(0.00026)	(0.00021)
Age	−0.00002	−0.00001	0.000004	−0.000004
	(0.00001)	(0.00001)	(0.00001)	(0.00001)
Arrest Charge				
Number of Charges	−0.000002	−0.00004	−0.000005	0.00003
	(0.00001)	(0.00003)	(0.00003)	(0.00007)
Any Drug Charge	−0.00022	−0.00008	−0.00012	−0.00008
	(0.00028)	(0.00024)	(0.00031)	(0.00023)
Any Violent Crime Charge	−0.00043	0.00001	0.00038	−0.00013
	(0.00030)	(0.00018)	(0.00026)	(0.00017)
Any Property Charge	−0.00038	−0.00038	0.00023	−0.00070
	(0.00027)	(0.00028)	(0.00029)	(0.00035)
Any DUI Charge	0.00047	0.00049	0.00100	0.00012
	(0.00057)	(0.00030)	(0.00093)	(0.00042)
Defendant Priors				
Prior FTA	−0.00014	−0.00005	0.00012	−0.00003
	(0.00023)	(0.00020)	(0.00024)	(0.00015)
Prior Misdemeanor Arrest	0.00024	−0.00012	0.00009	0.00010
	(0.00025)	(0.00017)	(0.00028)	(0.00018)
Prior Felony Arrest	−0.00007	−0.000005	−0.00043	0.00040
	(0.00036)	(0.00023)	(0.00032)	(0.00022)
Prior Violent Felony Arrest	−0.00042	0.00012	−0.00001	−0.00020
	(0.00029)	(0.00021)	(0.00025)	(0.00018)
Prior Misdemeanor Conviction	−0.00009	0.00050	0.00042	−0.00013
	(0.00030)	(0.00021)	(0.00027)	(0.00017)
Prior Felony Conviction	0.00010	0.00024	−0.00040	−0.00041
	(0.00034)	(0.00023)	(0.00025)	(0.00019)
Prior Violent Felony Conviction	0.00040	−0.00084	−0.00004	0.0000001
	(0.00036)	(0.00030)	(0.00028)	(0.00024)
Joint p-value	0.05623	0.27401	0.24607	0.24712
Court × Time FE	✓	✓	✓	✓
Cases	99,463	171,241	112,517	186,035

Notes: This table reports OLS estimates for regressions of the constructed judge leniency measure on various defendant and case characteristics. These regressions are estimated separately over subsamples defined on the race of the defendant and whether the defendant was charged with a felony offense. Standard errors, reported in parentheses, are clustered at the defendant and judge level. The joint p-value is based on the F-statistic for whether all defendant and case characteristics are jointly significant. See Section 5.3 of the main text for discussion.