Identifying Prediction Mistakes in Observational Data*

Ashesh Rambachan†

October 3, 2022

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? Uncovering systematic prediction mistakes is difficult as the preferences and information sets of decision makers are unknown to researchers. In this paper, I characterize behavioral and econometric assumptions under which systematic prediction mistakes can be identified in observational empirical settings such as hiring, medical testing, and pretrial release. I derive a statistical test for whether the decision maker makes systematic prediction mistakes under these assumptions, and provide methods for conducting inference on the ways in which the decision maker's predictions are systematically biased. As an empirical illustration, I analyze the pretrial release decisions of judges in New York City, estimating that at least 20% of judges make systematic prediction mistakes about failure to appear risk given defendant characteristics. Motivated by this behavioral analysis, I estimate the effects of replacing judges with algorithmic decision rules, and find that automating decisions where systematic prediction mistakes occur weakly dominates the status quo.

*First version: June 2021. I am especially grateful to Isaiah Andrews, Sendhil Mullainathan, Neil Shephard, Elie Tamer and Jens Ludwig for their invaluable feedback and advice. I thank Nano Barahona, Laura Blattner, Iavor Bojinov, Raj Chetty, Bo Cowgill, Will Dobbie, Xavier Gabaix, Matthew Gentzkow, Ed Glaeser, Peter Hull, Larry Katz, Daniel Martin, Maria Polyakova, Jonathan Roth, Joshua Schwartzstein, Jesse Shapiro, and seminar participants at Princeton, Microsoft Research, Harvard Business School, MIT, Northwestern Kellogg, University of Pennsylvania, Berkeley, Columbia, Yale, Carnegie Mellon, Stanford, the ACM SIGecom Winter Meeting, the Chamberlain Online Seminar, and the NBER Economics of AI Conference 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.

†Microsoft Research New England and MIT: ashesh.a.rambachan@gmail.com

1 Introduction

Decision makers, such as doctors, judges, and managers, must often make consequential decisions based on predictions of unknown outcomes. For example, in deciding whether to detain a defendant awaiting trial, a judge predicts what the defendant will do if released 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 (e.g., Meehl, 1954; Tversky and Kahneman, 1974) 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] In assessing whether such machine learning based models can improve decision-making, empirical researchers attempt to evaluate decision makers' implicit predictions through comparisons of their choices against those made by predictive models.[2]

Yet uncovering systematic prediction mistakes from decisions 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 diverge from the model not because she is making systematic prediction mistakes, but rather she 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*a*; 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 such an econometric framework for analyzing whether a decision maker makes systematic prediction mistakes and to characterize how their predictions are systematically biased. This clarifies what can (and cannot) be identified about systematic prediction mistakes from data and empirically relevant assumptions about behavior, and maps those assumptions into

[1] Risk assessment tools are used in criminal justice systems throughout the United States (Stevenson, 2018; Albright, 2019; Dobbie and Yang, 2019; Stevenson and Doleac, 2019; Yang and Dobbie, 2020). Clinical risk assessments aid doctors in diagnostic and treatment decisions (Obermeyer and Emanuel, 2016; Beaulieu-Jones et al., 2019; Abaluck et al., 2020; Chen et al., 2020). For applications in consumer finance, see for example Einav, Jenkins and Levin (2013), Fuster et al. (2022), Gillis (2019), Dobbie et al. (2021), and Blattner and Nelson (2021). For discussions of workforce analytics and resume screening software, see Autor and Scarborough (2008), Hoffman, Kahn and Li (2018), Li, Raymond and Bergman (2020), Raghavan et al. (2020), and Frankel (2021).

[2] See, for example, Kleinberg et al. (2015), Chalfin et al. (2016), Chouldechova et al. (2018), Hoffman, Kahn and Li (2018), Kleinberg et al. (2018*a*), Erel et al. (2019), Ribers and Ullrich (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). See Camerer (2019) for a recent review of this literature.

statistical inferences about systematic prediction mistakes.

I consider empirical settings, such as pretrial release, medical diagnosis, and hiring, in which a decision maker must make decisions for many individuals based on a prediction of some unknown outcome using each individual's characteristics. These characteristics are observable to both the decision maker and the researcher. The available data on the decision maker's choices and associated outcomes suffer from 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).

This paper then makes four main contributions. First, I characterize behavioral and econometric assumptions under which systematic prediction mistakes can be identified in these empirical settings. Second, under these assumptions, I provide a complete empirical characterization of whether the decision maker's choices reflect systematic prediction mistakes, and show how researchers can statistically test whether these conditions are satisfied. Third, I provide methods for conducting inference on the ways in which the decision maker's predictions are systematically biased. These contributions enable empirical researchers to answer a wide array of behavioral questions under weak assumptions. Finally, I apply this econometric framework to analyze the pretrial release decisions of judges in New York City as an empirical illustration.

I explore the 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 beliefs about the outcome given the characteristics as well as some 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 given the characteristics to lie in this identified set, what I call "accurate beliefs." If there exists no utility function in a researcher-specified class nor any distribution of 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 a sharp empirical characterization of expected utility maximization at accurate beliefs over an economically rich, benchmark class of utility functions. Using this characterization, I show that systematic prediction mistakes are *untestable* without further assumptions. If either all characteristics of individuals directly affect the decision maker's utility function or the missing data can take any value, then any variation in the decision maker's conditional choice probabilities can be rationalized. However, placing an exclusion restriction on which characteristics may directly affect the decision maker's utility function and constructing informative bounds on the missing data restores the testability of expected utility maximization behavior. Under such an exclusion restriction, variation in the decision maker's choices across characteristics that do not directly affect the utility function must only arise due to variation in beliefs. The decision maker's

beliefs given the characteristics and her private information must further be Bayes-plausible with respect to some distribution of the outcome given the characteristics that lies in the identified set. Together this implies testable restrictions on the decision maker's choices across characteristics that do not directly affect utility. Behavioral assumptions about the decision maker's utility function and econometric assumptions to address the missing data problem are therefore sufficient to identify systematic prediction mistakes. Testable restrictions arise from the *joint* null hypothesis that the decision maker maximizes expected utility at accurate beliefs and that their utility function satisfies the conjectured exclusion restriction.

With this framework in place, I further establish that the data are informative about the magnitudes of the decision maker's systematic prediction mistakes. I extend the behavioral model to only require that the decision maker's approximately maximize expected utility, meaning that they are only within some expected utility cost of being optimal. This is a computational device to summarize the extent to which the decision maker's choices deviate from expected utility maximization at accurate beliefs, but takes no stand on what drives the decision maker's misoptimizations. I sharply characterize the identified set of expected utility costs implied by the decision maker's choices. Using this characterization, I then show the total expected utility cost to the decision maker of their systematic prediction mistakes is the optimal value of a linear program, and the share of systematic prediction mistakes in their decisions is the optimal value of a mixed-integer linear program.

Finally, I explore one particular mechanism for the decision maker's misoptimization by analyzing whether the data are informative about the ways in which the decision maker's beliefs are systematically biased. To do so, I allow the decision maker to have possibly inaccurate beliefs about the unknown outcome and sharply characterize the identified set of utility functions at which the decision maker's choices are consistent with "inaccurate" expected utility maximization.[3] This takes no stand on the behavioral foundations for the decision maker's inaccurate beliefs, and so it encompasses various frictions or mental gaps such as inattention to characteristics or representativeness heuristics (e.g., Sims, 2003; Handel and Schwartzstein, 2018; Gabaix, 2019). I derive bounds on an interpretable parameter that summarizes the extent to which the decision maker's beliefs overreact or underreact to the characteristics of individuals. For a fixed pair of characteristic values, these bounds summarize whether the decision maker's beliefs about the outcome vary more ("overreact") or less than ("underreact") the true conditional distribution of the outcome across these values.

As an empirical illustration, I analyze the pretrial release 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.[4] For each judge, I observe the conditional probability that she releases

[3]The decision maker's beliefs about the outcome conditional on the characteristics are no longer required to lie in the identified set for the conditional distribution of the outcome given the characteristics.

[4]Several empirical papers also study the NYC pretrial release system. Leslie and Pope (2017) estimates the effects

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 detained defendants 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. We cannot logically rule out that the judge's release decisions reflect either a utility function that varies richly based on defendant characteristics or sufficiently predictive private information absent further assumptions.

However, empirical researchers often assume that while judges may engage in taste-based discrimination on a defendant's race, other defendant characteristics such as prior pretrial misconduct history only affects 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 exclusion restrictions and quasi-experimental bounds on the missing data, my identification results establish that expected utility maximization behavior 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 detained defendants? If not, there is no combination of a utility function that satisfies 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 a range of exclusion restrictions and quasi-experimental bounds on the failure to appear rate among detained defendants, 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. I further find that these systematic prediction mistakes 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 therefore driven by release decisions on defendants at the tails of the predicted risk distribution.

of pretrial detention on criminal case outcomes. Arnold, Dobbie and Hull (2022) and Arnold, Dobbie and Hull (2020) estimate whether judges and pretrial risk assessments respectively discriminate against black defendants. Kleinberg et al. (2018*a*) studies whether a machine learning-based risk assessment could improve pretrial outcomes in New York City.

Finally, to highlight policy lessons from this behavioral analysis, I explore the implications of replacing decision makers with algorithmic decision rules in the New York City pretrial release setting. Since supervised machine learning methods are tailored to deliver accurate predictions (Mullainathan and Spiess, 2017; Athey, 2017), such algorithmic decision rules may improve outcomes by correcting systematic prediction mistakes. I estimate the effects of replacing judges who were found to make systematic prediction mistakes with an algorithmic decision rule. Automating decisions 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, which is measured as a weighted average of the failure to appear rate among released defendants and the pretrial detention rate. Automating decisions whenever the human decision maker makes systematic prediction mistakes can therefore be a free lunch. Fully replacing judges with the algorithmic decision rule, however, has ambiguous effects that depend on the parametrization of social welfare. In fact, for some parametrizations of social welfare, I find that fully automating decisions can lead to up to 25% reductions in worst-case expected social welfare relative to the judges' observed decisions.

This paper relates to a growing empirical literature that evaluates decision makers' predictions through either comparisons of their choices against those made by machine learning based models (e.g., Kleinberg et al., 2018*a*; Mullainathan and Obermeyer, 2022) or estimating parametric, structural models of decision making behavior (e.g., Abaluck et al., 2016; Arnold, Dobbie and Hull, 2022; Chan, Gentzkow and Yu, 2022). The econometric framework in this paper only requires the researcher to specify an exclusion restriction on which characteristics affect the decision maker's utility function. I otherwise flexibly model the decision maker's utility function, allowing it to vary arbitrarily across non-excluded characteristics. I also model the decision maker's information environment fully nonparametrically. This enables researchers to both identify and characterize systematic prediction mistakes in many empirical settings under weaker assumptions than existing research.

Most closely related is Kleinberg et al. (2018*a*) which directly compares the pretrial release decisions of all judges in New York City against an estimated, machine learning based decision rule. Viewed through the lens of my identification analysis, by comparing the pooled choices of judges against an estimated machine learning based decision rule, Kleinberg et al. (2018*a*) is limited to making statements about 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. In contrast, I conduct my analysis judge-by-judge, allow each judge's utility function to flexibly vary based on defendant characteristics, allow for unrestricted heterogeneity in utility functions across judges, and allow private information to vary arbitrarily across judges. I further characterize the magni-

tudes of judges' systematic prediction mistakes, and the ways in which their beliefs about failure to appear risk are systematically biased.

The econometric framework in this paper builds 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 so far due to the difficulty of collecting SDSC data (Gabaix, 2019; Rehbeck, 2020). I focus on common empirical settings in which the data suffer from a missing data problem, and show that these settings can approximate ideal SDSC data by using quasi-experimental variation to address the missing data problem. Martin and Marx (2021) study the identification of taste-based discrimination by a decision maker in a binary choice experiment, providing bounds on the decision maker's group-dependent threshold rule. 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.

My identification analysis follows in the spirit of the information design literature (e.g., Kamenica and Gentzkow, 2011; Bergemann and Morris, 2016, 2019) by asking whether there exists *any* private information such that the decision maker's choices are consistent with expected utility maximization behavior. Several recent papers take this approach in different settings or to answer different questions. Syrgkanis, Tamer and Ziani (2018) studies auctions, and Magnolfi and Roncoroni (2021) studies entry games. Bergemann, Brooks and Morris (2019) bound the welfare changes of counterfactuals that alter unknown information structures in both both single-agent and multiplayer settings. Gualdani and Sinha (2020) also analyzes single-agent, discrete-choice settings under weak assumptions on the decision maker's information environment, whereas I focus on directly testing whether choices are consistent with expected utility maximization behavior and characterizing the uncovered violations.

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 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 $y^* := (y_1^*, \ldots, y_K^*) \in \mathcal{Y} \subseteq [0, 1]^K$. 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 throughout that 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, however, a *missing data* or *selective labels* problem: we only observe Y^* if the decision maker selected $C = 1$ (Rubin, 1976; Kleinberg et al., 2018*a*). 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

$$\pi_c(x) := P(C = c \mid X = x) \text{ for all } c \in \{0, 1\} \text{ and } x \in \mathcal{X}, \tag{1}$$

and the observable conditional outcome probabilities are

$$P_1(y^* \mid x) := P(Y^* = y^* \mid C = 1, X = x) \text{ for all } x \in \mathcal{X}. \tag{2}$$

The conditional outcome probabilities $P_0(y^* \mid x) := P(Y^* = y^* \mid C = 0, X = x)$ is not identified due to the missing data problem.[5] The true outcome probabilities $P(y^* \mid x) := P(Y^* = y^* \mid X = x)$ is 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.[6]

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*a*; Arnold, Dobbie and Hull, 2022). The outcome $Y^* = Y_1^* \in \{0, 1\}$ 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 current charges filed against the defendant, and the defendant's prior arrest/conviction record. We observe the characteristics of each defendant, whether the judge released them, and whether the defendant failed to appear in court if the judge released them. 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 observed. The conditional failure to appear rate among detained defendants $P_0(y^* \mid x)$ is unobserved. ▲

Example (Medical Testing and Diagnosis)**.** A doctor decides whether to conduct a costly medical test or make a particular diagnosis (e.g., Abaluck et al., 2016; Ribers and Ullrich, 2019; Chan, Gentzkow and Yu, 2022). For example, shortly after an emergency room visit, a doctor decides

[5] I adopt the convention that $P(Y^* = y^* \mid C = c, X = x) = 0$ if $\pi_c(x) = 0$.

[6] Screening decisions are a leading class of "prediction policy problems" (Kleinberg et al., 2015). Other examples include loan approvals (e.g., Fuster et al., 2022; Dobbie et al., 2021), academic admissions (Dawes, 1979; Kleinberg et al., 2018*b*), child welfare screenings (Chouldechova et al., 2018), and disability insurance screenings (e.g., Benitez-Silva, Buchinsky and Rust, 2004; Low and Pistaferri, 2015, 2019).

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 $Y^* = Y_1^* \in \{0, 1\}$ 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. We observe the characteristics of each patient, whether the doctor conducted a stress test, and whether the patient had a heart attack if the doctor conducted a stress test. 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 observed. The conditional heart attack rate among untested patients $P_0(y^* \mid x)$ is unobserved. ▲

Example (Hiring)**.** A hiring manager decides whether to hire job applicants $C \in \{0, 1\}$ (Autor and Scarborough, 2008; Chalfin et al., 2016; Hoffman, Kahn and Li, 2018; Frankel, 2021).[7] 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. We observe the characteristics of each applicant, whether the manager hired the applicant, and their on-the-job productivity if hired. The manager's conditional hiring rate $\pi_1(x)$ and the conditional distribution of tenure lengths among hired applicants $P_1(y^* \mid x)$ are observed. The distribution of tenure lengths among rejected applicants $P_0(y^* \mid x)$ is unobserved. ▲

The main text makes two simplifying assumptions for exposition: (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. Appendix C analyzes a more general setting in which the decision maker faces a treatment assignment problem with multiple choices, which nests the main text as a special case.

Finally, for a finite set $\mathcal{A}$, let $\Delta(\mathcal{A})$ denote the set of all probability distributions on $\mathcal{A}$. 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 $X = x$, and 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 assumptions about the missing data problem in the form of bounds on the unknown conditional outcome probabilities.

Assumption 1. For each $x \in \mathcal{X}$, there exists a known subset $\mathcal{B}_x \subseteq \Delta(\mathcal{Y})$ satisfying $P_0(\cdot \mid x) \in \mathcal{B}_x$. The collection of all bounds is $\mathcal{B} = \{\mathcal{B}_x \colon x \in \mathcal{X}\}$.

[7] The setting also applies to job interview decisions (Cowgill, 2018; Li, Raymond and Bergman, 2020), where the choice $C \in \{0, 1\}$ is whether to interview an applicant, and $Y^* = Y_1^* \in \{0, 1\}$ is whether the applicant is hired by the firm.

In some cases, we may analyze the decision maker's choices without placing any further assumptions on the missing data, which corresponds to setting $\mathcal{B}_x$ equal to the set of all conditional outcome probabilities. In other cases, researchers may use quasi-experimental variation or introduce additional structural assumptions to provide informative bounds on the unknown conditional outcome probabilities, as I discuss in Section 3.

Under Assumption 1, the joint distribution $(X, C, Y^*) \sim P(\cdot)$ is partially identified. The sharp identified set for the true outcome probabilities given $x \in \mathcal{X}$, denoted $\mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_x)$, equals the set of $\widetilde{P}(\cdot \mid x) \in \Delta(\mathcal{Y})$ satisfying

$$\widetilde{P}(y^* \mid x) = \widetilde{P}_0(y^* \mid x)\pi_0(x) + P_1(y^* \mid x)\pi_1(x) \tag{3}$$

for all $y^* \in \mathcal{Y}$ and some $\widetilde{P}_0(\cdot \mid x) \in \mathcal{B}_x$.

2.3 Behavioral model

In this setting, I examine the restrictions placed on the decision maker's choices by expected utility maximization at accurate beliefs. Under this model, the decision maker's information set for each individual consists of their characteristics and some additional private information. 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 are often not recorded. I place no distributional assumptions on the decision maker's private information.[8] The decision maker forms beliefs about the unknown outcome based on this information set and selects a choice to maximize expected utility.

Suppose the researcher partitions the characteristics $x := (x_0, x_1)$ with $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_1$. The expected utility maximization model is summarized by a utility function and a joint distribution over the characteristics, private information, choices and outcomes, denoted $(X, C, V, Y^*) \sim Q$, that satisfies three conditions.

Definition 1. A *utility function* $u\colon \{0,1\} \times \mathcal{Y} \times \mathcal{X}_0 \to \mathbb{R}$ specifies the payoff associated with each choice-outcome pair at characteristics $x_0 \in \mathcal{X}_0$. Let $\mathcal{U}$ denote the feasible set of utility functions specified by the researcher.

Definition 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 $(X, V, C, Y^*) \sim Q$ satisfying

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

[8] This contrasts with structural models of decision-making behavior that assume the decision maker's information set is summarized by some known parametric distribution. See, for example, Abaluck et al. (2016); Chan, Gentzkow and Yu (2022) in medical diagnosis and Arnold, Dobbie and Hull (2022) in pretrial release.

$Q(c \mid x, v) > 0$,

$$\mathbb{E}_Q\left[u(c, Y^*; X_0) \mid X = x, V = v\right] \geq \mathbb{E}_Q\left[u(c', Y^*; X_0) \mid X = x, V = v\right].$$

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

iii. Data Consistency: For all $x \in \mathcal{X}$, there exists $\widetilde{P}_0(\cdot \mid x) \in \mathcal{B}_{0,x}$ 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 $y^* \in \mathcal{Y}$.

The *identified set of utility functions*, denoted $\mathcal{H}_P(u; \mathcal{B}) \subseteq \mathcal{U}$, is the set of utility functions $u \in \mathcal{U}$ such that there exists $(X, V, C, Y^*) \sim Q$ satisfying (i)-(iii).

The decision maker's choices are consistent with expected utility maximization if three conditions are satisfied. First, if choice c is selected with positive probability given (X, V) under the model Q, then it must have been optimal to do so in an expected utility sense ("Expected Utility Maximization"). The decision maker may flexibly randomize across choices whenever they are indifferent. Second, the decision maker's choices must be independent of the outcome given the characteristics and private information under the model Q ("Information Set"), formalizing the sense in which the decision maker's information set consists of only (X, V). Finally, the joint distribution of characteristics, choices and outcomes under the model Q must be consistent with the observable joint distribution P ("Data Consistency").[9]

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_0 directly affect the decision maker's utility function. In medical testing and diagnosis, researchers assume that a doctor's payoffs are constant across patients, and patient characteristics affect beliefs about the probability of an underlying medical condition (e.g., Abaluck et al., 2016; Chan, Gentzkow and Yu, 2022; Mullainathan and Obermeyer, 2022). In

[9]The expected utility maximization model relates to recent developments on Roy-style selection (Mourifie, Henry and Meango, 2019; Henry, Meango and Mourifie, 2020) and marginal outcome tests for taste-based discrimination (Canay, Mogstad and Mountjoy, 2020; Hull, 2021). Defining the expected benefit functions $\Lambda_c(x, v) = \mathbb{E}_Q\left[U(c, Y^*; X_0) \mid X = x, V = v\right]$ for $c \in \{0, 1\}$, the expected utility maximization model is a generalized Roy model that assume $X_0 \in \mathcal{X}_0$ affects both the utility function and beliefs, whereas $X_1 \in \mathcal{X}$ and private information $V \in \mathcal{V}$ only affect beliefs. The expected utility maximization model is an "incomplete" model since it makes no assumptions on how indifferences are resolved (e.g., Tamer, 2003).

pretrial release, the utility function specifies a judge's relative payoffs from detaining a defendant that would not fail to appear in court and releasing a defendant that would fail to appear in court. Researchers often assume these payoffs may vary based on only some defendant characteristics. For example, judges may engage in tasted-based discrimination against black defendants (Becker, 1957; Arnold, Dobbie and Yang, 2018; Arnold, Dobbie and Hull, 2022), be more lenient towards younger defendants (Stevenson and Doleac, 2019), or be more harsh towards defendants charged with violent crimes (Kleinberg et al., 2018*a*).

Since this is a substantive economic assumption, I discuss three ways to specify such exclusion restrictions on the decision maker's utility function. First, as mentioned, such utility exclusion restrictions are common in empirical research. The researcher may therefore appeal to established modelling choices to guide this assumption. Second, the exclusion restriction may be normatively motivated, summarizing social or legal restrictions on what characteristics ought not to directly enter the decision maker's utility function. Third, 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.

2.3.2 Accurate beliefs and systematic prediction mistakes

If Definition 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 $P(\cdot \mid x)$ as a consequence of Data Consistency.

Lemma 2.1. *If the decision maker's choices are consistent with expected utility maximization, then* $Q(\cdot \mid x) \in \mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_x)$ *for all* $x \in \mathcal{X}$.

The decision maker's implied beliefs $Q(\cdot \mid x)$ are therefore *accurate* in this sense if their choices are consistent with expected utility maximization. Conversely, 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 accurate beliefs in the identified set for the true outcome probability. The decision maker in this case is acting as-if their implied beliefs given the characteristics are systematically mistaken.

Definition 3. The decision maker is making *systematic prediction mistakes* based on the characteristics if their choices are inconsistent with expected utility maximization, meaning $\mathcal{H}_P(u; \mathcal{B}) = \emptyset$.

The interpretation of a systematic prediction mistake is tied to both the researcher-specified bounds on the missing data $\mathcal{B}$ (Assumption 1) and the feasible set of utility functions $\mathcal{U}$ (Definition 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.[10] 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 3 must therefore be interpreted as a systematic prediction mistake that can be identified given 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 the testable implications of expected utility maximization at accurate beliefs under various assumptions on the decision maker's utility function and the missing data. Over a benchmark class of utility functions, identifying systematic prediction mistakes requires both behavioral assumptions on which characteristics may directly affect the decision maker's utility function and econometric assumptions that generate informative bounds on the unobservable conditional outcome probabilities. Under these conditions, testing whether the decision maker's choices are consistent with expected utility maximization at accurate beliefs is equivalent to testing many moment inequalities.

3.1 Characterization result

I derive conditions under which the decision maker's choices are consistent with expected utility maximization over the class of *linear* utility functions.

Definition 4. The class of *linear utility functions* is the set of utility functions satisfying $u(c, y^*; x_0) = \sum_{k=1}^{K} u_{1,k}(x_0) y_k^* c + u_{0,k}(x_0)(1-y_k^*)(1-c)$, where $u_{1,k}(x_0), u_{0,k}(x_0) \leq 0$, $|u_{1,k}(x_0) + u_{0,k}(x_0)| = 1$ for all $x_0 \in \mathcal{X}_0$.

This is an economically rich class that captures many common empirical intuitions. The parameters $u_{1,k}(x_0), u_{0,k}(x_0) \leq 0$ summarize the cost of ex-post errors for each outcome – 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 characteristics X_0 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, this assumes it is costly for the judge to detain a defendant that would not fail to appear or release a defendant

[10] Consider an extreme case in which $P(\cdot \mid x)$ is partially identified under bounds $\mathcal{B}_x$ but point identified under alternative bounds $\widetilde{\mathcal{B}}_x$. Under Definitions 2-3, systematic prediction mistakes at bounds $\widetilde{\mathcal{B}}_x$ means that the decision maker's implied beliefs $Q(\cdot \mid x)$ do not equal the point identified quantity $P(\cdot \mid x)$, yet systematic prediction mistakes at bounds $\mathcal{B}_x$ means that the decision maker's implied beliefs $Q(\cdot \mid x)$ do not lie in the identified set $\mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_x)$.

that would fail to appear, but places no restrictions on how these costs vary across defendant characteristics X_0, 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 $u_{1,1}(x_0)/u_{2,1}(x_0)$. The class of linear utility functions is therefore a useful benchmark to understand the testable implications of expected utility maximization at accurate beliefs.

For $x_0 \in \mathcal{X}_0$, define $\Pi_1(x_0) := \{x_1 \in \mathcal{X}_1 \colon \pi_1(x_0, x_1) > 0\}$ and $\Pi_0(x_0) := \{x_1 \in \mathcal{X}_1 \colon \pi_0(x_0, x_1) > 0\}$. 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 $\bar{\mu}_0(x) := \max_{\tilde{P}(\cdot|x) \in \mathcal{B}_x} \mu_0(x)$.

Theorem 3.1. *The decision maker's choices are consistent with expected utility maximization at some linear utility function if and only if, for all $x_0 \in \mathcal{X}_0$,*

$$\max_{x_1 \in \Pi_1(x_0)} \mu_1(x_0, x_1) \leq \min_{x_1 \in \Pi_0(x_0)} \bar{\mu}_0(x_0, x_1) \tag{4}$$

Otherwise, $\mathcal{H}_P(u; \mathcal{B}) = \emptyset$, and the decision maker is making systematic prediction mistakes.

Corollary 3.1. *The identified set of linear utility functions $\mathcal{H}_P(u; \mathcal{B})$ equals the set of all utility functions satisfying Definition 4 and, for all $x_0 \in \mathcal{X}_0$,*

$$\max_{x_1 \in \Pi_1(x_0)} \mu_1(x_0, x_1) \leq \sum_{k=1}^K |u_{0,k}(x_0)| \leq \min_{x_1 \in \Pi_0(x_0)} \bar{\mu}_0(x_0, x_1). \tag{5}$$

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 their posterior expectation for $\bar{Y}^*$. The threshold may vary across characteristics X_0, and it is incomplete since it takes no stand on how possible indifferences are resolved. The main step in the proof of Theorem 3.1 shows that the conditional outcome probabilities summarize all possible posterior beliefs that could arise by applying Bayes rule to any distribution of private information and accurate beliefs. A threshold rule on their posterior beliefs is therefore observationally equivalent to a threshold rule on these conditional outcome probabilities. The researcher's assumptions about the missing data therefore restrict the decision maker's private information and implied beliefs. The inequalities (4) then check whether any value of the conditional outcome probabilities consistent with the researcher's bounds (Assumption 1) could reproduce the decision maker's choices under such a threshold rule.[11]

[11] 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. The unknown outcome and characteristics can be interpreted as a payoff-relevant state-of-the-world, and each choice is a state-dependent lottery over payoffs $u(c, y^*; x_0)$. Due to the missing data problem and because the decision maker's true payoffs are unknown, Theorem 3.1 searches over all possible beliefs and payoffs.

In Appendix C.2, I provide a complete empirical characterization of expected utility maximization behavior for more general treatment assignment problems. Theorem 3.1 then applies this general characterization over the class of linear utility functions in screening decisions. The key insight underlying this general characterization is that the decision maker's choices are consistent with expected utility maximization if and only if a hypothetical information designer could induce a decision maker with accurate beliefs to take the observed choices by providing additional information to them via some information structure (e.g., Bergemann and Morris, 2019; Kamenica, 2019). The information structure is the decision maker's private information under the expected utility maximization model. I must simultaneously check whether the information designer could induce the observed choices at *any* accurate beliefs $\tilde{P}(\cdot \mid x) \in \mathcal{H}(P(\cdot \mid x); \mathcal{B}_x)$ due to the missing data problem, and *any* utility function $u \in \mathcal{U}$ since the decision maker's true payoffs are unknown.[12]

3.2 When are systematic prediction mistakes identifiable?

If the inequalities in Theorem 3.1 are violated, there exists no linear utility function, private information, nor accurate beliefs at which the decision maker's choices are consistent with expected utility maximization. By examining cases in which these inequalities are always satisfied, I characterize leading cases in which we cannot identify systematic prediction mistakes in the decision maker's choices under our stated assumptions.

Corollary 3.2. *The decision maker's choices are always consistent with expected utility maximization at accurate beliefs and some linear utility function if either:*

(i) all characteristics directly affect utility (i.e, $\mathcal{X} = \mathcal{X}_0$) and $\mu_1(x_0) \leq \bar{\mu}_0(x_0)$ for all $x_0 \in \mathcal{X}_0$;

(ii) $\bar{\mu}_0(x) = K$ for all $x \in \mathcal{X}$.

Corollary 3.2(i) highlights the necessity of placing an exclusion restriction on which characteristics directly affect the decision maker's utility function. If all characteristics directly affect the decision maker's utility function (i.e., $\mathcal{X} = \mathcal{X}_0$), then the decision maker's choices are consistent with expected utility maximization whenever our missing data assumptions are compatible with the decision maker observing useful private information. More precisely, if the conditional expectation of $\bar{Y}^*$ given $C = 0$ can always 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 X_0 can always be constructed that rationalizes the

[12]The identification problem underlying expected utility maximization therefore relates to a recent literature on robust information design with unknown prior beliefs (e.g., Kosterina, 2022) and unknown utility functions (e.g., Babichenko et al., 2021). The decision maker's initial prior beliefs and payoffs are both unknown in my identification analysis, whereas this literature studies optimal persuasion mechanisms when only one is unknown at a time.

decision maker's choices. If our missing data assumptions allow for the decision maker to observe useful private information in this weak sense, then an exclusion restriction on the decision maker's utility function is necessary. Unfortunately, Corollary 3.2(ii) also establishes that imposing such an exclusion restriction alone may be insufficient to restore the identifiability of expected utility maximization. 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. Identifying systematic prediction mistakes over the class of linear utility functions therefore requires *both* 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.[13]

Under such assumptions Theorem 3.1 provides interpretable conditions for identifying systematic prediction mistakes. At any fixed $x_0 \in \mathcal{X}_0$, does there exist some $x_1 \in \mathcal{X}_1$ 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 $x_1' \in \mathcal{X}_1$? If so, the decision maker could do strictly better by raising their probability of selecting choice $C = 0$ at x_1' and lowering their probability of selecting choice $C = 1$ at x_1 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 discrimination 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. Theorem 3.1 shows that these *misrankings* completely 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 exclusion restriction.

[13] In comparing an algorithmic decision rule against a decision maker, Kleinberg et al. (2018*a*) introduce "omitted payoffs bias" to refer to the concern that the algorithm may predict an outcome different than the outcome the decision maker bases their choices on. Corollary 3.2 highlights that even if we correctly specify the outcome Y^*, we still cannot identify systematic prediction mistakes over the class of linear utility functions without utility exclusion restrictions and informative bounds on the unobservable conditional outcome probabilities.

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*a*; Arnold, Dobbie and Hull, 2022), and doctors may be randomly assigned to patients in medical testing (Abaluck et al., 2016; Chan, Gentzkow and Yu, 2022).[14]

Assumption 2 (Random Instrument)**.** Let $Z \in \mathcal{Z}$ be a finite support instrument. The joint distribution $(X, Z, C, Y^*) \sim P(\cdot)$ satisfies $(X, Y^*) \perp\!\!\!\perp Z$, 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 2, denoting its sharp identified set as $\mathcal{H}_P(\mu_0(x, z))$. In the case where the instrument arises through the random assignment of decision makers, $\mathcal{H}_P(\mu_0(x, z))$ corresponds to sharp bounds on the conditional outcome probabilities for for a single decision maker.

Proposition 3.1. *Suppose Assumption 2 holds. For any $(x, z) \in \mathcal{X} \times \mathcal{Z}$ with $\pi_0(x, z) > 0$, $\mathcal{H}_P(\mu_0(x, z)) = [\underline{\mu}_0(x, z), \overline{\mu}_1(x, z)]$, where*

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

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

Proposition 3.1 follows from worst-case bounds on $\mu(x, z) := \mathbb{E}[\bar{Y}^* \mid X = x]$ (e.g., Manski, 1989, 1994) and point identification of $\mu_1(x, z), \pi_0(x, z)$. Appendix E.1 extends these bounds to allow for the instrument to be quasi-randomly assigned, which will be used in the empirical application to pretrial release decisions in New York City.

Furthermore, under the expected utility maximization model, Assumption 2 only requires that the decision maker's initial beliefs given the characteristics do not depend on the instrument but places no other behavioral restrictions.

Proposition 3.2. *Suppose Assumption 2 holds. If the decision maker's choices are consistent with expected utility maximization at some utility function u and joint distribution $(X, Z, V, C, Y^*) \sim Q$, then $Y^* \perp\!\!\!\perp Z \mid X$ under Q.*

[14] Other examples of instruments appear in empirical research. Mullainathan and Obermeyer (2022) argue that there is quasi-random, day-of-week variation in the likelihood doctors conduct stress tests for a heart attack due to staffing constraints. The introduction of or changes to recommended decision-making guidelines may also affect decision makers' choices (e.g., Albright, 2019; Abaluck et al., 2020).

This is an immediate consequence of Definition 2. Requiring that the decision maker's implied beliefs be accurate imposes that the randomly assigned instrument cannot affect their beliefs about the outcome given the characteristics. Both utility functions and private information can 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.[15,16] These bounds also require no form of parametric extrapolation across decision makers (e.g., Arnold, Dobbie and Hull, 2022), which may be sensitive to functional form assumptions when there is no "extremely lenient" value of the instrument (i.e., $\pi_0(x,z) \approx 0$).

Using the instrumental variable bounds $\mathcal{H}_P(\mu_0(x,\tilde{z}))$, we 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. The characterization reduces to a system of many moment inequalities.

Proposition 3.3. *Suppose Assumption 2 holds, 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* $x_0 \in \mathcal{X}_0$*, pairs* $x_1, \tilde{x}_1 \in \mathcal{X}_1$ *and* $\tilde{z} \in \mathcal{Z}$*,*

$$\mu_1(x_0, x_1, z) - \overline{\mu}_{0,\tilde{z}}(x_0, \tilde{x}_1, z) \leq 0, \tag{6}$$

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

The number of moment inequalities (6) equals $|\mathcal{X}_0| \cdot |\mathcal{X}_1|^2 \cdot (|\mathcal{Z}| - 1)$, 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.[17] I return to this problem and discuss my practical solution below in the empirical application to the NYC pretrial system.

[15] de Chaisemartin (2017); Frandsen, Lefgren and Leslie (2019) analyze violations of monotonicity in settings where decision makers are randomly assigned to individuals. Allowing private information to vary across decision makers allows for rich variation in "skill" (Chan, Gentzkow and Yu, 2022).

[16] Lakkaraju et al. (2017) and Kleinberg et al. (2018*a*) use the random assignment of decision makers to evaluate an algorithmic decision rule $\tilde{C}$ by imputing $P(Y^* = 1 \mid \tilde{C} = 1)$. In contrast, Proposition 3.1 constructs bounds on a decision maker's conditional expectation $\mu_0(x,z)$.

[17] While there are high-dimensional moment inequality procedures such as Chernozhukov, Chetverikov and Kato (2019) and Bai, Santos and Shaikh (2021), these require the sample analogues of the moments to be expressed as averages of i.i.d observations and so are not directly applicable.

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. By extending the expected utility maximization model, I next show that researchers can further investigate the magnitudes of the decision maker's systematic prediction mistakes and the ways in which the decision maker's beliefs are systematically biased. First, I show that misrankings in the decision maker's choices characterize the expected utility cost and the share of systematic prediction mistakes in their decisions. Second, I derive bounds on the extent to which the decision maker's beliefs overreact or underreact to variation in the characteristics.

4.1 Identifying the costs and share of systematic prediction mistakes

To characterize the expected utility cost and share of systematic prediction mistakes in the decision maker's choices, I weaken Definition 2 to only require that the decision maker's choices *approximately* maximize expected utility, meaning they are within some $\epsilon(x) \geq 0$ of being optimal.

Definition 5. The decision maker's choices *approximately maximize* expected utility at accurate beliefs if there exists a utility function $u \in \mathcal{U}$, expected utility costs $\epsilon(x) \geq 0$ for all $x \in \mathcal{X}$, and joint distribution $(X, V, C, Y^*) \sim Q$ 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 $Q(c \mid x, v) > 0$,

$$\mathbb{E}_Q\left[u(c, Y^*; X_0) \mid X = x, V = v\right] \geq \mathbb{E}_Q\left[u(c', Y^*; X_0) \mid X = x, V = v\right] - \epsilon(x).$$

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

I use approximate expected utility maximization as a computational device to characterize the extent to which the decision maker's choices deviate from expected utility maximization at accurate beliefs.[18] First, notice the decision maker's choices are always trivially consistent with approximate expected utility maximization for large enough expected utility costs $\epsilon(x) \geq 0$ (i.e., $\mathcal{H}_P(\epsilon; \mathcal{B}) = \emptyset$). Second, if the decision maker's choices are consistent with expected utility maximization at accurate beliefs (Definition 2), then the decision maker's choices are consistent with

[18] My approach therefore relates to recent work measuring violations of utility maximization behavior in consumer demand settings (e.g., Apesteguia and Ballester, 2015; Allen and Rehbeck, 2020; Echenique, Saito and Imai, 2021).

approximate expected utility maximization at $\epsilon(x) = 0$ for all $x \in \mathcal{X}$. Therefore, at each characteristic $x \in \mathcal{X}$, the smallest $\epsilon(x)$ satisfying Definition 5 summarizes how large are the violations of expected utility maximization at accurate beliefs implied by the decision maker's choices.

The identified set of expected utility costs $\epsilon(x) \geq 0$ over the class of linear utility functions is sharply characterized by misrankings in the decision maker's choices.

Theorem 4.1. *Assume $0 < \pi_1(x) < 1$ 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 $x = (x_0, x_1)$, $x' = (x_0, x_1')$,*

$$\mu_1(x) - \overline{\mu}_0(x') - \epsilon(x) - \epsilon(x') \leq 0. \tag{7}$$

The sharp identified set of expected utility costs $\mathcal{H}_P(\epsilon; \mathcal{B})$ equals the set of all $\epsilon = \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ satisfying (7).

Appendix C.3 provides a complete empirical characterization of approximate expected utility maximization behavior for more general treatment assignment problems, and Theorem 4.1 applies this general characterization over the class of linear utility functions in screening decisions. This misrankings characterization of expected utility costs implies tractable characterizations of the total expected utility cost and the share of systematic prediction mistakes in the decision maker's choices.

4.1.1 Bounding the expected utility costs of systematic prediction mistakes

The lower bound on the expected utility cost of systematic prediction mistakes to the decision maker is

$$\underline{\mathcal{E}} := \min_{\epsilon} \sum_{x \in \mathcal{X}} P(x)\epsilon(x) \text{ s.t. } \epsilon \in \mathcal{H}_P(\epsilon; \mathcal{B}). \tag{8}$$

$\underline{\mathcal{E}}$ summarizes how worse off is the decision maker in an expected utility sense relative to hypothetical choices that correctly optimized expected utility given her information set. Notice $\underline{\mathcal{E}} = 0$ if and only if the decision maker's choices are consistent with expected utility maximization at accurate beliefs by construction. Theorem 4.1 implies $\underline{\mathcal{E}}$ can be equivalently characterized as the optimal value of the linear program

$$\begin{aligned} \underline{\mathcal{E}} = &\min_{\epsilon} \sum_{x \in \mathcal{X}} P(x)\epsilon(x) \\ &\text{s.t. } \epsilon(x) \geq 0 \text{ for all } x \in \mathcal{X}, \\ &\quad \mu_1(x) - \overline{\mu}_0(x') - \epsilon(x) - \epsilon(x') \leq 0 \text{ for all pairs } x = (x_0, x_1), x' = (x_0, x_1'). \end{aligned} \tag{9}$$

This linear program is feasible since the constraints are always satisfied, and so the optimal value of its sample analogue $\widehat{\underline{\mathcal{E}}}$ converges to the population value by Theorem 3.5 of Shapiro (1991). Misrankings therefore recover the sharp lower bound on the expected utility cost of systematic prediction mistakes to the decision maker. This lower bound applies to any linear utility function, accurate beliefs, and distribution of private information.

Since $\underline{\mathcal{E}}$ is expressed in expected utility units, its magnitude may nonetheless be difficult to interpret directly. In Appendix E.2, I show, for a scalar outcome $Y^* = Y_1^*$, $\underline{\mathcal{E}}$ can be translated into an equivalent fraction of ex-post errors that arose from the decision maker's systematic prediction mistakes. Given an optimal solution to (9), we can recover the decision maker's implied linear utility function $u_{0,1}(x_0), u_{1,1}(x_0)$, and calculate an equivalent reduction in ex-post errors $\mathbb{E}[Y^* \cdot C]$ that would produce the same expected utility cost $\underline{\mathcal{E}}$. In the pretrial release example, $\underline{\mathcal{E}}$ is the judge's total expected utility cost of their systematic prediction mistakes about failure to appear risk. Using the judge's implied costs of releasing a defendant that would fail to appear in court, $\underline{\mathcal{E}}$ can be translated into an equivalent reduction in the fraction of defendants that are released and fail to appear that would produce the same 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. I define a subset of characteristics $\mathcal{X}_R \subseteq \mathcal{X}$ to be *rationalizable at accurate beliefs* if there exists a utility function $u \in \mathcal{U}$ and joint distribution $(X, V, C, Y^*) \sim Q$ satisfying Definition 2 only over $x \in \mathcal{X}_R$. The largest rationalizable subset $\overline{\mathcal{X}}_R$ of characteristics is defined as

$$\overline{\mathcal{X}}_R := \arg\max_{\mathcal{X}_R \subseteq \mathcal{X}} \sum_{x \in \mathcal{X}_R} P(x) \text{ s.t. } \mathcal{X}_R \text{ is rationalizable at accurate beliefs,} \tag{10}$$

and the share of rationalizable decisions is $P(\overline{\mathcal{X}}_R) := \sum_{x \in \overline{\mathcal{X}}_R} P(x)$. That is, $\overline{\mathcal{X}}_R$ is the largest subset of decisions for which there exists some utility function, accurate beliefs, and private information that could rationalize the decision maker's choices. Notice if the decision maker's choices are consistent with expected utility maximization at accurate beliefs, then $\overline{\mathcal{X}}_R = \mathcal{X}$ and $P(\overline{\mathcal{X}}_R) = 1$. Furthermore, by definition, the decision maker's choices are not rationalizable at accurate beliefs for any $\overline{\mathcal{X}}_R \cup \tilde{\mathcal{X}}$ with $\tilde{\mathcal{X}} \subseteq \mathcal{X} - \mathcal{X}_R$. The *share of systematic prediction mistakes* in the decision maker's choices is therefore given by $1 - P(\overline{\mathcal{X}}_R)$.

Theorem 4.1 implies that the share of systematic prediction mistakes can be equivalently characterized by the optimal value of the following optimization program.

Theorem 4.2. *The share of systematic prediction mistakes in the decision maker's choices satisfies*

$1 - P(\overline{\mathcal{X}}_R) = \underline{\lambda}$, *where*

$$\begin{aligned} \underline{\lambda} := & \min_{\epsilon} \sum_{x \in \mathcal{X}} P(x) 1\{\epsilon(x) > 0\} \\ & \textit{s.t. } \epsilon(x) \geq 0 \textit{ for all } x \in \mathcal{X}, \\ & \quad \mu_1(x) - \overline{\mu}_0(x') - \epsilon(x) - \epsilon(x') \leq 0 \textit{ for all pairs } x = (x_0, x_1), x' = (x_0, x_1'). \end{aligned} \tag{11}$$

Appendix E.3 further shows that (11) is equivalent to a mixed-integer linear program, which can be solved accurately using modern optimization solvers. The decision maker's misrankings therefore additionally recover the share of systematic prediction mistakes in the decision maker's choices.

4.2 Bounding inaccurate beliefs based on characteristics

Not only can we summarize the magnitudes of the decision maker's systematic prediction mistakes, I next show that we can further investigate the ways in which the decision maker's beliefs are systematically biased. Since the definition of expected utility maximization required that the decision maker act as-if their beliefs were accurate (Lemma 2.1), misrankings may indicate that the decision maker's beliefs are *inaccurate* – that is, their implied beliefs do not lie in the identified set $\mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_x)$. This is a common behavioral hypothesis in empirical applications. Empirical researchers conjecture that judges may systematically mis-predict failure to appear risk based on defendant characteristics, and the same concern arises in analyses of medical decisions.[19]

To investigate whether the decision maker's choices maximize expected utility at inaccurate beliefs, I modify "Data Consistency" in Definition 2.

Definition 6. 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 $(X, V, C, Y^*) \sim Q$ satisfying (i) Expected Utility Maximization, (ii) Information Set as in Definition 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 $\widetilde{P}(y^* \mid x) = P_1(y^* \mid x)\pi_1(x) + \widetilde{P}_0(y^* \mid x)\pi_0(x)$.

[19] Kleinberg et al. (2018*a*) write, "a primary source of error is that all quintiles of judges misuse the signal available in defendant characteristics available in our data" (pg. 282-283). In the medical treatment setting, Currie and Macleod (2017) write, "we are concerned with doctors, who for a variety of possible reasons, do not make the best use of the publicly available information at their disposal to make good decisions" (pg. 5).

Definition 6 drops the restriction that the decision maker's implied beliefs must lie in the identified set for the true outcome probabilities. It only requires that the joint distribution $(X, V, C, Y^*) \sim Q$ under the model matches the joint distribution of the observable data $(X, C, Y) \sim P$ if the decision maker's model-implied beliefs, $Q(\cdot \mid x)$, are replaced with some true outcome probabilities in the identified set, $\widetilde{P}(\cdot \mid x) \in \mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_x)$. Since it places no direct restrictions on the decision maker's implied prior beliefs $Q(\cdot \mid x)$, behavior consistent with expected utility maximization at inaccurate beliefs could arise from various behavioral mechanisms.[20]

The next result characterizes whether the decision maker's observed choices are consistent with expected utility maximization at inaccurate beliefs and some linear utility function.

Theorem 4.3. *Assume $\widetilde{P}(\cdot \mid x) > 0$ for all $\widetilde{P}(\cdot \mid x) \in \mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_x)$, $x \in \mathcal{X}$ 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 $\widetilde{P}_0(\cdot \mid x) \in \mathcal{B}_{0,x}$ and non-negative weights $\omega(y^*; x)$ satisfying, for all $x_0 \in \mathcal{X}_0$,*

$$\max_{\tilde{x}_1 \in \Pi_1(x_0)} \mathbb{E}_{\widetilde{P}}[\omega_1(Y^*; X)\bar{Y}^* \mid C = 1, X = (x_0, \tilde{x}_1)] \leq \min_{\tilde{x}_1 \in \mathcal{P}_0(x_0)} \mathbb{E}_{\widetilde{P}}[\omega_0(Y^*; X)\bar{Y}^* \mid C = 0, X = (x_0, \tilde{x}_1)]$$

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 under $(X, C, Y^*) \sim \widetilde{P}(\cdot)$ defined as

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

The weights $\omega(y^*; x)$ are the likelihood ratio of the decision maker's implied beliefs relative to some conditional distribution of the outcomes given the characteristics in the identified set. Since the decision maker's prediction mistakes only arise from misspecification of beliefs $Q_{Y^*}(\cdot \mid x)$, her posterior beliefs under the model are proportional to the likelihood ratio between her beliefs and the underlying outcome distribution. Since expected utility maximization over the class of linear utility functions is equivalent to a threshold rule on the decision maker's posterior expectation for $\bar{Y}^*$, expected utility maximization at inaccurate beliefs and some linear utility function is therefore equivalent to a threshold rule on this reweighed conditional expectation. Theorem 4.3 shows that these misrankings at the reweighed conditional expectations completely characterize expected utility maximization at inaccurate beliefs and some linear utility function.

[20] For example, Definition 6 is consistent with decision maker's implied beliefs being inaccurate due to inattention to the characteristics (e.g., Sims, 2003; Gabaix, 2014; Caplin and Dean, 2015) or use of representativeness heuristics (e.g., Gennaioli and Shleifer, 2010; Bordalo et al., 2016; Bordalo, Gennaioli and Shleifer, 2021). Developing formal tests for these specific behavioral mechanisms in observational settings is beyond the scope of this paper.

4.2.1 Bounding inaccurate beliefs for a binary outcome

For a screening decision with a binary outcome $Y^* = Y_1^* \in \{0, 1\}$, Theorem 4.3 implies a bound on the extent to which the decision maker's beliefs given the characteristics overreact or underreact to variation in the characteristics. As a first step, the same intuition underlying Theorem 4.3 can be used to bound the decision maker's reweighted utility function $\omega(y^*; x)u(c, y^*; x_0)$ in a screening decision with a binary outcome.

Theorem 4.4. *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 $\widetilde{P}(\cdot \mid x) \in \mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_x)$ 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 $x \in \mathcal{X}$,*

$$P_1(1 \mid x) \leq \frac{\omega(0; x)u_{0,1}(x_0)}{\omega(0; x)u_{0,1}(x_0) + \omega(1; x)u_{1,1}(x_0)} \leq \overline{P}_0(1 \mid x), \tag{12}$$

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 6.

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 $\tau(x) := \frac{\omega(0;x)u_{0,0}(x_0)}{\omega(0;x)u_{0,0}(x_0)+\omega(1;x)u_{1,1}(x_0)}$ to be the decision maker's reweighed utility threshold. If the reweighed utility threshold were known, then the decision maker's implied prediction mistake $\delta(x)$ could be backed out.

Corollary 4.1. *Under the same conditions as Theorem 4.4, for any $x_0 \in \mathcal{X}_0$ and $x_1, x_1' \in \mathcal{X}_1$,*

$$\frac{(1 - \tau(x_0, x_1))/\tau(x_0, x_1)}{(1 - \tau(x_0, x_1'))/\tau(x_0, x_1')} = \frac{\delta(x_0, x_1)}{\delta(x_0, x_1')}. \tag{13}$$

The ratio $\delta(x_0, x_1)/\delta(x_0, x_1')$ summarizes the extent to which the decision maker's beliefs overreact or underreact to variation in the characteristics relative to the true conditional distribution. By definition, if $\delta(x_0, x_1)/\delta(x_0, x_1')$ is less than one, then the decision maker's beliefs about the relative probability of $Y_1^* = 1$ versus $Y_1^* = 0$ (i.e., $Q(1 \mid x)/Q(0 \mid x)$) varies less across the characteristics (x_0, x_1) and (x_0, x_1') than the true outcome probabilities. The decision maker's implied beliefs therefore *underreact* across these characteristics. Analogously if $\delta(x_0, x_1)/\delta(x_0, x_1')$ is strictly greater than one, then the decision maker's implied beliefs *overreact* across the characteristics in this relative sense.[21] Since Theorem 4.4 provides an identified set for the reweighted

[21]The parameter $\frac{\delta(x_0,x_1)}{\delta(x_0,x_1')}$ summarizes how relative changes in the decision maker's beliefs compare to relative changes in the true outcome probabilities. As an example, suppose $\widetilde{P}(1 \mid x_0, x_1) = 4/5, \widetilde{P}(1 \mid x_0, x_1') = 1/5$ and

utility thresholds, an identified set for the implied prediction mistake $\delta(x_0, x_1)/\delta(x_0, x_1')$ can in turn be constructed by computing the ratio (13) for each pair $\tau(x_0, x_1), \tau(x_0, x_1')$ satisfying (12).

These bounds on the extent to which the decision maker's beliefs overreact or underreact are obtained only by assuming that the decision maker's linear utility function satisfy the conjectured exclusion restriction. Under such an exclusion restriction, variation in the decision maker's choices across excluded characteristics must only arise due to variation in beliefs. Examining how conditional outcome probabilities vary across characteristics relative to the decision maker's choices is therefore informative about the decision maker's systematic prediction mistakes. For this reason, utility exclusion restrictions are sufficient to partially identify the extent to which variation in the decision maker's beliefs are biased.[22]

5 Do pretrial release judges make systematic prediction mistakes?

As an empirical illustration, I apply this econometric framework to analyze the pretrial release 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 exclusion restrictions on their utility functions, their pretrial release decisions are inconsistent with expected utility maximization at accurate beliefs about failure to appear risk given defendant characteristics. Rejections of expected utility maximization at accurate beliefs are driven primarily by release decisions on defendants at the tails of the predicted risk distribution, and arise because their implied beliefs systematically underreact to defendant characteristics.

5.1 Pretrial release decisions in New York City

I analyze the pretrial release system in New York City, which has been previously studied in Leslie and Pope (2017), Kleinberg et al. (2018*a*) and Arnold, Dobbie and Hull (2022). The NYC pretrial system is an ideal setting to apply this econometric framework for three reasons. First, as discussed in Kleinberg et al. (2018*a*), the NYC pretrial system narrowly asks judges to only consider failure to appear risk in deciding whether to release a defendant. I initially define the outcome $Y^* =$

the decision maker's beliefs are $Q(1 \mid x_0, x_1) = 2/3, Q(1 \mid x, x_1') = 1/3$. The true odds ratios are $\frac{\widetilde{P}(1|x_0,x_1)}{\widetilde{P}(0|x_0,x_1)} = 4$, $\frac{\widetilde{P}(1|x_0,x_1')}{\widetilde{P}(0|x_0,x_1')} = 1/4$. The decision maker's perceived odds ratio are $\frac{Q(1|x_0,x_1)}{Q(0|x_0,x_1)} = 2$, $\frac{Q(1|x_0,x_1')}{Q(0|x_0,x')} = 1/2$. The implied prediction mistake $\frac{\delta(x_0,x_1)}{\delta(x_0,x_1')}$ equals $1/4$. If instead $Q(1 \mid x_0, x_1) = 3/4, Q(1 \mid x_0, x_1') = 3/7$, then the decision maker's perceived odds ratios would equal 3, $3/4$ at characteristics (x_0, x_1), (x_0, x_1') respectively, but $\frac{\delta(x_0,x_1)}{\delta(x_0,x_1')}$ again equals $1/4$ even though the decision maker's beliefs differ.

[22] This result relates to Proposition 1 in Martin and Marx (2021), which shows that utilities and prior beliefs are not separately identified in state-dependent stochastic choice environments (see also Bohren et al. (2020)). These negative results arise because previous authors exclusively focused on settings in which there are no additional characteristics of decisions beyond those which affect both utility and beliefs.

$Y_1^* \in \{0, 1\}$ to be whether the defendant would fail to appear in court if released, and explore the robustness of my empirical findings to alternative outcomes.[23] Second, the NYC pretrial release system is one of the largest in the country, and consequently I observe many judges making a large number of pretrial release decisions. Finally, judges in New York City are quasi-randomly assigned to cases within court-by-time cells, which implies bounds on the conditional failure to appear rate among detained defendants.

I observe 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.[24] I apply additional sample restrictions to construct the main estimation sample, which consists of 569,256 cases heard by 265 unique judges.[25] I test whether each of the top 25 judges that heard the most cases make systematic prediction mistakes about failure to appear risk in their pretrial release decisions. These top 25 judges altogether heard 243,118 cases in the main estimation sample, and each judge heard at least 5,000 cases.

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 each defendant, and whether the assigned judge released or detained the defendant.[26] 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 Supplement Table S1 provides descriptive statistics about the main estimation sample and the cases heard by the top 25 judges. Overall, 72.0% of defendants are released in the main estimation sample, whereas 73.6% of defendants assigned to the top 25 judges were released. Defendants in the main estimation sample are similar on demographic information and current charge information to defendants assigned to the top 25 judges. However, defendants assigned to the top 25 judges have less extensive prior criminal records. Online Supplement Table S2 reports the same descriptive statistics broken out by whether the defendant was released or detained, revealing that judges appear to respond to defendant characteristics in their release decisions. Among defendants

[23] In Online Supplement H.3, I instead define the binary outcome to be whether the defendant would either fail to appear in court or be re-arrested, finding similar results.

[24] I construct the set of arrests subject to a pretrial release decision as in Kleinberg et al. (2018*a*), removing (i) desk appearance tickets, (ii) cases disposed at arraignment, and (iii) cases adjourned in contemplation of dismissal, and (iv) duplicate cases.

[25] 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.

[26] Judges in New York City decide whether to release a defendant without conditions ("release on recognizance"), require the defendant to post a chosen amount of bail, or deny bail altogether. Following Kleinberg et al. (2018*a*); Arnold, Dobbie and Hull (2022), I collapse these choices into the binary decision of whether to release or detain. In Online Supplement H.4 extends the empirical analysis and finds that at least 32% of judges make decisions that 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.

assigned to the top 25 judges, released and detained defendants differ demographically: 50.8% of released defendants are white and 19.7% are female, whereas only 40.7% of detained defendants are white and only 10.6% are female. Released and detained defendants also differ on their current charge and criminal record. For example, only 28.8% of defendants released by the top 25 judges face a felony charge, yet 58.6% of detained defendants face a felony charge.

I test whether the release decisions of 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. I test whether there exists misrankings in their decisions 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. 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

As mentioned earlier, a key practical challenge in testing whether judges' release decisions satisfy Proposition 3.3 is that the number of moment inequalities is large, and as a consequence the number of observations per characteristic cell is extremely small. For example, 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. Focusing only on the judge that heard the most cases over the sample period, there are on average only 1.87 cells per characteristic cell .

To deal with this practical challenge, I instead test whether there are implied misrankings in the judges' decisions over a coarsened partition of the characteristics. Formally, define $D\colon \mathcal{X} \to \{1, \ldots, N_d\}$ to be some partition of 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 $D(\cdot)$. Now let $\mu_c(x_0, d) := \mathbb{E}[\bar{Y}^* \mid C = c, X_0 = x_0, D(X) = d]$ and $\pi_c(x_0, d) := P(C = c \mid X_0 = x_0, D(X) = d)$ for $c \in \{0, 1\}$.

Proposition 5.1. *If the decision maker's choices are consistent with expected utility maximization at accurate beliefs and some linear utility function, then, for all* $x_0 \in \mathcal{X}_0$

$$\max_{d \in \mathcal{D}_1(x_0)} \mu_1(x_0, d) \leq \min_{x \in \mathcal{D}_0(x_0)} \overline{\mu}_0(x_0, d), \tag{14}$$

where $\mathcal{D}_1(x_0) := \{d\colon \pi_1(x_0, d) > 0\}$ *and* $\mathcal{D}_0(x_0) := \{d\colon \pi_0(x_0, d) > 0\}$.

Provided $N_d \ll |\mathcal{X}_1|$, the number of moment inequalities implied (14) is drastically reduced,

and can be tested using standard moment inequality methods that rely on an asymptotic normal approximation to the sample moments (Canay and Shaikh, 2017; Molinari, 2020). WSearching 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.

The choice of $D(\cdot)$ is clearly crucial for the power of this falsification test to detect violations of expected utility maximization at accurate beliefs. I argue that a natural choice is to construct $D(\cdot)$ using supervised machine learning methods that predict the outcome $\bar{Y}^*$ on the pretrial release decisions of other judges. Given an estimated prediction function $\hat{f}\colon \mathcal{X} \to [0, K]$, $D(\cdot)$ can be defined by binning the characteristics X into percentiles of predicted risk within each $x_0 \in \mathcal{X}_0$.[27] 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_1 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.

Proposition 5.2. *Assume $\hat{f}(x) = \mu(x) := \mathbb{E}[\bar{Y}^* \mid X = x]$, $D(x)$ is defined as the level sets of $\hat{f}(x)$, and that $\mu(X)$ is sufficient for the decision maker's private information and tie-breaking rule, meaning $V \mid \{Y^*, X_0, X_1\} \sim V \mid \{Y^*, X_0, \mu(X)\}$ and $C \mid \{V, X_0, X_1\} \sim C \mid \{V, X_0, \mu(X)\}$ under Q. Then the decision maker's choices are consistent with expected utility maximization at some linear utility function if and only if, for all $x_0 \in \mathcal{X}_0$, (14) is satisfied.*

Proposition 5.2 provides a novel connection between out-of-sample prediction and identification in analyzing systematic prediction mistakes under the additional behavioral assumption that the excluded characteristics X_1 only enter the decision maker's information set through their initial beliefs. Under this behavioral assumption, the excluded characteristics X_1 only affect the decision maker's beliefs through the true conditional expectation $\mu^*(x)$, and so the extent to which the inequalities in (14) are non-sharp is therefore driven by how well the estimated prediction function recovers $\mu^*(x)$.

In Appendix D, I show that the earlier 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 retain intuitive interpretations after this coarsening step. Over the coarsening, we can recover a lower bound on the worst-case expected utility cost and worst-case share of systematic prediction mistakes respectively. The implied prediction mistake across values $D(X) = d, D(X) = d'$, denoted by $\delta(x_0, d)/\delta(x_0, d')$, measures how the decision maker's implied beliefs of their own ex-post mistakes vary relative to the true probability of ex-post mistakes across

[27] There may already exist a benchmark risk score. In pretrial release systems, the widely-used Public Safety Assessment summarizes observable defendant characteristics into an integer-valued risk score (e.g., Stevenson, 2018; Albright, 2019). In medicine, commonly used risk assessments summarize observable patient characteristics into an integer-valued risk score (e.g., Obermeyer and Emanuel, 2016; Lakkaraju and Rudin, 2017).

values $D(X) = d, D(X) = d'$.

In my empirical analysis of the NYC pretrial system, I apply these results by constructing a partition of the characteristics $X \in \mathcal{X}$. I predict failure to appear risk among defendants released by all other judges within each value of the payoff-relevant characteristics $x_0 \in \mathcal{X}_0$, 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 $x_0 \in \mathcal{X}_0$. The prediction function is an ensemble that averages the predictions of an elastic net model and a random forest.[28] Over defendants released by the top 25 judges, the ensemble model achieves an area under the receiver operating characteristic (ROC) curve, or AUC, of 0.693 when the payoff-relevant characteristics are defined as race-by-age cells and an AUC of 0.694 when the payoff relevant characteristics are defined as race-by-felony charge cells. Both achieve similar performance on released black and white defendants.

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.[29] 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 payoff-relevant characteristic cell (defined by race-by-age cells or race-by-felony-charge cells), reporting the coefficient estimates in Online Supplement Tables S3-S4. 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 pooled main estimation sample and for all subsamples, except for young black defendants.

I use the quasi-random assignment of judges to construct bounds on the failure to appear rate among defendants detained by each judge in the top 25.[30] 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 Appendix E.1, the bound on the failure to appear rate among defendants with $X_0 = x_0, D(X) = d$ for a particular judge using leniency

[28] I use three-fold cross-validation to tune the penalties for the elastic net model. The random forest is constructed using the `R` package `ranger` at the default hyperparameter values (Wright and Ziegler, 2017).

[29] There are two relevant features that suggest judges are as-if randomly assigned to cases in NYC. First, bail judges are assigned to shifts in each of the five county courthouses in NYC based on a rotation calendar system. Second, there is limited scope for public defenders or prosecutors to influence which judge will decide any particular case. See Kleinberg et al. (2018*a*); Arnold, Dobbie and Hull (2022) for further discussion.

[30] Online Supplement H.2 alternatively assumes that the failure to appear rate among detained defendants can be no greater than some chosen constant times the observed failure to appear rate among released defendants. I find qualitatively similar results using this alternative bounding strategy.

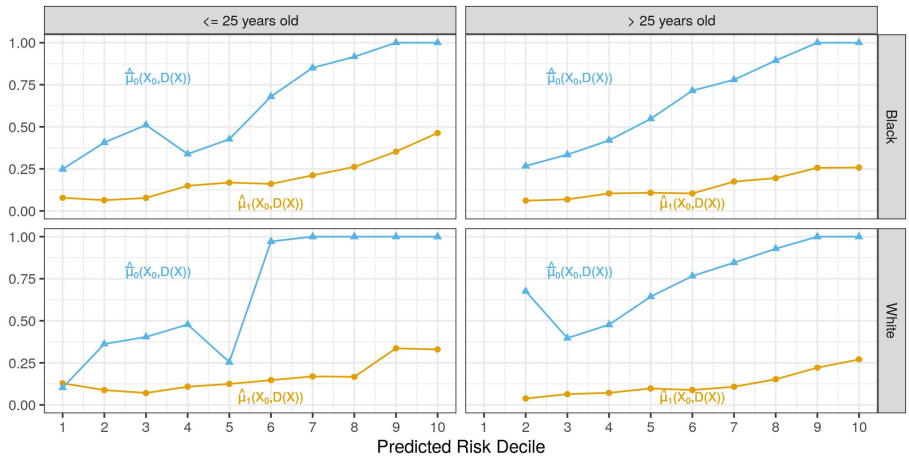
quintile $\tilde{z} \in \mathcal{Z}$ depends on quantities $\mathbb{E}[P(C = 1, Y_1^* = 1 \mid X_0 = x_0, D(X) = d, Z = \tilde{z}, T) \mid X_0 = x_0, D(X) = d]$ and $\mathbb{E}[P(C = 0 \mid X_0 = x_0, D(X) = d, Z = \tilde{z}, T) \mid X_0 = x_0, D(X) = d]$, where $T \in \mathcal{T}$ denotes the court-by-time cells and the expectation averages over all cases assigned to this particular judge. I model the conditional probabilities as

$$1\{C = 1, Y_1^* = 1\} = \sum_{x_0,d,z} \beta_{x_0,d,z}^{c,y_1^*} 1\{X_0 = x_0, D(X) = d, Z = z\} + \phi_t + \epsilon \tag{15}$$

$$1\{C = 0\} = \sum_{x_0,d,z} \beta_{x_0,d,z}^{c} 1\{X_0 = x_0, D(X) = d, Z = z\} + \phi_t + \nu, \tag{16}$$

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 $\hat{\beta}_{x_0,d,\tilde{z}}^{c}, \hat{\beta}_{x_0,d,\tilde{z}}^{c,y^*}$ to the average of the respective fixed effects associated with cases heard by the judge within each (x_0, d)-cell.

Figure 1: Failure to appear rate among released defendants and bound on the failure to appear rate among detained defendants by race-and-age cells for one judge in New York City.



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 1 plots the failure to appear rate among defendants released by the judge that heard the most cases and the 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 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). Figure A1 plots for each race-by-felony cell.[31]

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 1 for each judge in the top 25, I test whether there exist misrankings in their release decisions across deciles of predicted failure to appear risk (Proposition 5.1). The number of true rejections of these inequalities provides a lower bound on the number of judges whose choices are inconsistent with the joint null hypothesis that they are maximizing expected utility at accurate beliefs about failure to appear risk and some linear utility functions satisfying the conjectured exclusion restriction.

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. I estimate the variance-covariance matrix of the failure to appear rates among released defendants and upper bounds on the failure to appear rate among detained defendants using the bootstrap conditional on the payoff-relevant characteristics X_0, predicted risk decile $D(X)$ and leniency instrument Z. I use the conditional least-favorable hybrid test developed in Andrews, Roth and Pakes (2019) since it is computationally fast given the estimated moments and the variance-covariance matrix, and has desirable power properties (Rambachan and Roth, 2020).

Table 1: Estimated fraction of judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs about failure to appear risk given defendant characteristics.

	Utility Functions $u(c, y^*; x_0)$			
	No Characteristics	Race	Race + Age	Race + Felony Charge
Unadjusted Rejection Rate	48%	48%	48%	56%
Adjusted Rejection Rate	24%	24%	20%	32%

Notes: This table summarizes the results for testing whether there exists misrankings in the release decisions of each judge in the top 25 at linear utility functions $u(c, y^*; x_0)$ 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. The unadjusted rejection rate reports the fraction of judges in the top 25 whose pretrial release decisions violate the inequalities in Proposition 5.1 at the 5% level. 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.

Table 1 summarizes the results from testing whether there exists misrankings in the release decisions of each judge in the top 25 under various exclusion restrictions. After a multiple hy-

[31] Online Supplement H.2 reports findings using an alternative empirical strategy that bounds the conditional failure to appear rate among detained defendants using the observed failure to appear rate among released defendants.

pothesis 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. This is interpreted as a 95% lower bound on the number of true rejections of the misranking inequalities among the top 25 judges.[32] When both race and age are allowed to directly affect judges' preferences, 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 characteristics, allows for unrestricted heterogeneity in utility functions across judges, and allows for private information to vary arbitrarily across judges.

5.5 Bounding prediction mistakes based on defendant characteristics

Given that a large fraction of judges make pretrial release decisions that are inconsistent with expected utility maximization at accurate beliefs about failure to appear risk, I next investigate the ways in which their beliefs 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.

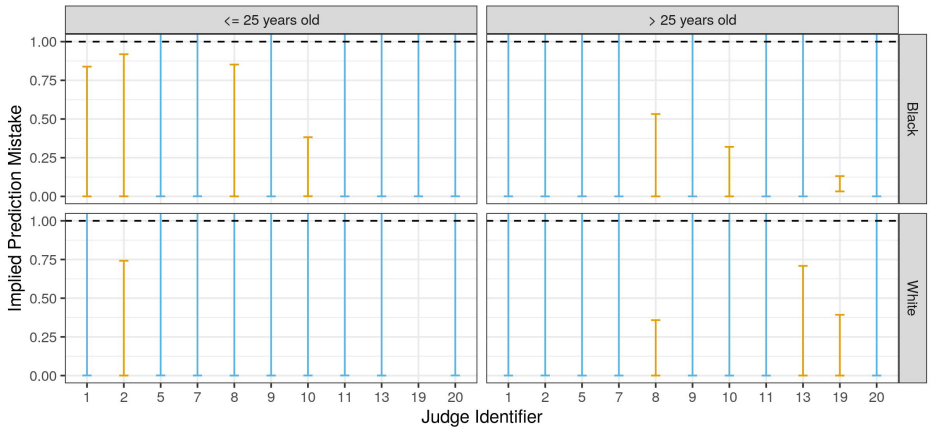
For each judge whose choices exhibit misrankings, I construct a 95% confidence interval for their implied prediction mistakes $\delta(x_0, d)/\delta(x_0, d')$ between the top decile d and bottom decile d' of the predicted failure to appear risk distribution. To do so, I first construct a 95% joint confidence set for the reweighted utility thresholds $\tau(x_0, d), \tau(x_0, d')$ at the bottom and top deciles of the predicted failure to appear risk distribution using test inversion based on Theorem 4.4, and then calculate $\frac{(1-\tau(x_0,d))/\tau(x_0,d)}{(1-\tau(x_0,d'))/\tau(x_0,d')}$ for each pair $\tau(x_0, d), \tau(x_0, d')$ in the joint confidence set as in Corollary 4.1.

Figure 2 plots the constructed confidence intervals for the implied prediction mistakes $\delta(x_0, d)/\delta(x_0, d')$ for each judge over the race-and-age cells. Figure A2 reports for race-and-felony charge cells.[33] 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 (Handel and Schwartzstein, 2018; Gabaix, 2019). While

[32] For m null hypotheses, let k be the number of false null hypotheses and let $\hat{k}$ be the number of rejections of a procedure controlling the family-wise error rate at the α-level. Observe $P(\hat{k} \leq k) = 1 - P(\hat{k} > k)$. Since $\{\hat{k} > k\} \subseteq \{\text{at least one false rejection}\}$, $P(\hat{k} > k) \leq P(\text{at least one false rejection})$, implying $P(\hat{k} \leq k) \geq 1 - P(\text{at least one false rejection}) \geq 1 - \alpha$.

[33] Online Supplement H shows that judges' implied beliefs underreact to variation in the latent outcome using alternative bounds on the missing data and alternatively defining the latent outcome to be any pretrial misconduct.

Figure 2: Estimated bounds on implied prediction mistakes between lowest and highest predicted failure to appear risk deciles made by judges within each race-by-age cell.



Notes: This figure plots the 95% confidence interval on the implied prediction mistake $\delta(x_0, d)/\delta(x_0, d')$ 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 violated the implied revealed preference inequalities (Table 1) 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 4.2.1 for theoretical details on the implied prediction mistake and Section 5.5 for the estimation details.

suggestive, developing formal tests for these specific behavioral mechanisms is beyond this paper's scope.

5.6 Which decisions violate expected utility maximization?

As a final step to investigate why the release decisions of judges in New York City are inconsistent with expected utility maximization at accurate beliefs, 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 2 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. These empirical findings together highlight that systematic prediction mistakes primarily occur on defendants at the tails of the predicted risk distribution.

Table 2: Location of the largest misranking among judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs about failure to appear risk given defendant characteristics.

	Utility Functions $u(c, y; x_0)$	
	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 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 judge's whose release decision violate the revealed preference inequalities at the 5% level, 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.

6 The welfare effects of algorithmic decision-making

I finally illustrate that this econometric analysis of systematic prediction mistakes has policy implications for the design of algorithmic decision systems by analyzing policy counterfactuals that replace judges with algorithmic decision rules in the NYC pretrial system.

For a binary outcome $Y^* = Y_1^* \in \{0, 1\}$, consider a policymaker whose payoffs are summarized by the linear social welfare function $u_{1,1}^* y_1^* c + u_{0,1}^*(1 - y_1^*)(1 - c)$. The policymaker evaluates a candidate decision rule $\pi_1^*(x) \in [0, 1]$, which denotes the probability $C = 1$ is chosen given $X = x$ (e.g., an algorithmic release rule in the pretrial release setting). Expected social welfare of the decision rule at $x \in \mathcal{X}$ is

$$P(1 \mid x)\pi_1^*(x)u_{1,1}^* + P(0 \mid x)\pi_0^*(x)u_{0,0}^*, \tag{17}$$

and total expected social welfare further averages according to the marginal distribution of characteristics.[34] Due to the missing data problem, expected welfare under any candidate decision rule is partially identified. Appendix E.4 characterizes its sharp identified set as an interval with bounds

[34]Expected social welfare does not incorporate additional fairness considerations that have received much attention in an influential computer science literature (e.g., see Barocas and Selbst, 2016; Barocas, Hardt and Narayanan, 2019). My analysis could be directly extended to allow payoffs to vary across groups defined by the characteristics (Rambachan et al., 2021) or a penalty depending on the composition of individuals that receive $C = 1$ (Kleinberg et al., 2018*b*).

computed by linear programs, as well as the sharp identified set of expected welfare under the decision maker's observed choices.

I compare expected social welfare under the status quo decisions of judges in NYC against expected social welfare under counterfactual algorithmic decision rules. Consistent with the stated objectives of the NYC pretrial system, I define the binary outcome to be whether a defendant would fail to appear in court. The cost of detaining an individual that would not fail to appear in court is $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$ and the cost of releasing a defendant that would fail to appear in court is $u^*_{1,1} = -1/|1+\tilde{u}|$. I report results as the relative cost of detaining an individual that would not fail to appear in court $|\tilde{u}|$ varies (i.e., an "unnecessary detention"). For a particular value $|\tilde{u}|$, 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_0 and each decile of predicted failure to appear risk $D(X)$. The threshold varies based on the parametrization of the social welfare function. See Appendix E.5 for further details.

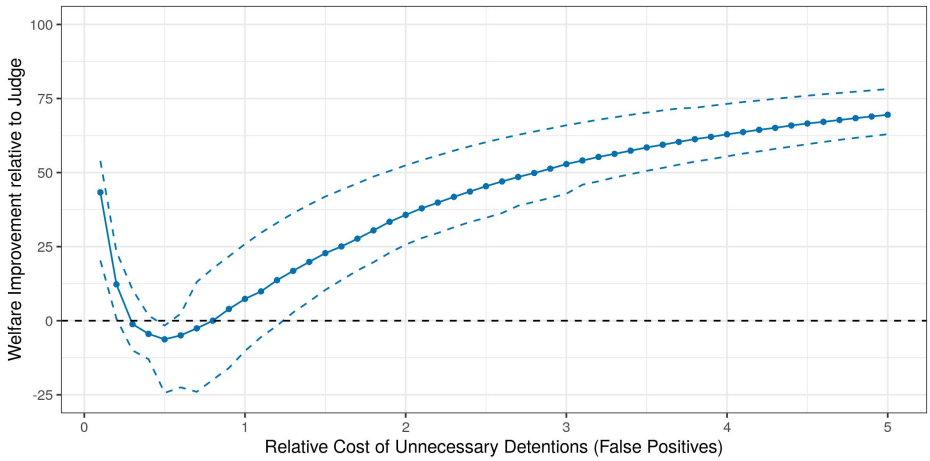
I construct 95% confidence intervals for expected social welfare under the algorithmic decision rule and the judge's observed released decisions, and report the ratio of worst-case expected social welfare under the algorithmic decision rule against the judge's observed release decisions. I conduct this exercise for each judge over the race-by-age cells, reporting the median, minimum and maximum gain across judges. Online Supplement H.1 reports results over the race-by-felony charge cells with similar findings.

6.1 Automating judges who make systematic prediction mistakes

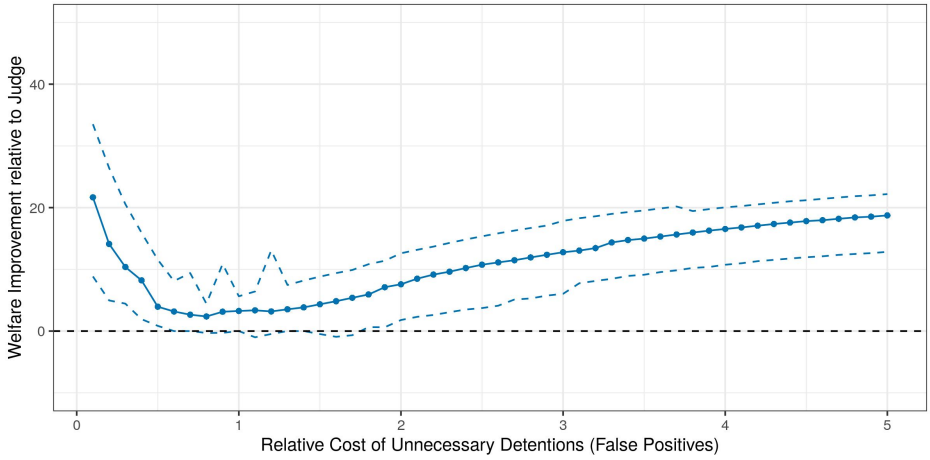
I first compare the the release decisions of judges who were found to make systematic prediction mistakes about failure to appear risk against an algorithmic decision rule that fully replaces ("automates") them over all defendants in Figure 3a. For most values of the social welfare function, worst-case expected social welfare under the full automation algorithmic decision rule is strictly larger than worst-case expected social welfare under these judges' decisions. Recall these judges primarily made systematic prediction mistakes over defendants in the tails of the predicted failure to appear risk distribution. Over the remaining defendants, however, their choices were found to be consistent with expected utility maximization at accurate beliefs about failure to appear risk. Consequently, the change in worst-case expected social welfare under the algorithmic decision rule is driven by three forces: first, the algorithmic decision rule may correct systematic prediction mistakes over the tails of the predicted risk distribution; second, the algorithmic decision rule corrects possible misalignment between the policymaker's objective and these judges' utility function over the remaining defendants; and third, these judges may observe useful private information over the remaining defendants that is unavailable to the algorithmic decision rule.

For social welfare costs of unnecessary detentions ranging over $|\tilde{u}| \in [0.3, 0.8]$, the algorithmic

Figure 3: Comparison of algorithmic decision rule relative to release decisions of judges that make systematic prediction mistakes about failure to appear risk.



(a) Full automation algorithmic decision rule



(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 that make systematic prediction mistakes about failure to appear risk. Panel (a) reports the comparison for an algorithmic decison 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 $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$. 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.

decision rule either leads to no improvement or strictly lowers worst-case expected total social welfare relative to these judges' decisions. Figure A3 plots the improvement in worst-case expected social welfare by the race of the defendant, highlighting that these costs are particularly large over white defendants. At these values, these judges may be sufficiently well-aligned with the policy-

maker and observe sufficiently predictive private information over the remaining defendants that it is costly to fully automate their decisions. To further investigate this, Figure A4 compares the release rates of the algorithmic decision rule against the observed release rates of these judges. The release rate of the algorithmic decision rule is most similar to the observed release rate of these judges precisely over the values of social welfare function where the status quo dominates the algorithmic decision rule.

The preceding behavioral analysis suggests that it would most valuable to automate these judges' decisions over defendants that lie in the tails of the predicted failure to appear risk distribution where they make systematic prediction mistakes. I next compare these judges' observed release decisions against an algorithmic decision rule that only automates decisions over defendants in the tails of the predicted failure to appear risk distribution and otherwise defers to the judges' observed decisions. This algorithmic decision rule therefore only corrects systematic prediction mistakes, and its welfare effects are plotted in Figure 3b. 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. This suggest that the identification of systematic prediction mistakes in a decision maker's choices provides a behavioral mechanism for recent machine learning methods that attempt to estimate whether a decision should be made by an algorithm or deferred to a decision maker (e.g., Madras, Pitassi and Zemel, 2018; Raghu et al., 2019; Wilder, Horvitz and Kamar, 2020). Deciding whether to automate or defer to an existing decision maker requires understanding whether the decision maker makes systematic prediction mistakes, and if so on what decisions.

6.2 Automating judges who do not make systematic prediction mistakes

Figure A5 reports the welfare effects of automating the release decisions of judges whose choices were found to be consistent with expected utility maximization at accurate beliefs. I find that automating these judges' release decisions may strictly lower worst-case expected social welfare for a range of social welfare costs of unnecessary detentions. It appears that these judges make pretrial release decisions as-if their utility functions were sufficiently aligned with the policymaker over these parametrizations of the social welfare function such that their private information leads to better decisions than the algorithmic decision rule. Figure A6 compares the release rates of the algorithmic decision rule against the observed release rates of these judges. Understanding the welfare effects of automating a decision maker whose decisions are consistent with expected maximization requires fully characterizing the tradeoff between the value of their private information against the degree to which they are misaligned with the policymaker, which is beyond the scope

of this paper.[35]

7 Conclusion

This paper develops an econometric framework for testing whether a decision maker makes systematic prediction mistakes in high stakes settings such as hiring, medical diagnosis and treatment, pretrial release, and many others. I characterized expected utility maximization behavior, where the decision maker maximizes some utility function at accurate beliefs about the outcome given the observable characteristics of each decision as well as some private information. I developed tractable statistical tests for whether the decision maker makes systematic prediction mistakes and methods for conducting inference on the ways in which their predictions are systematically biased. Analyzing the NYC pretrial release system, I found that a substantial fraction of judges make systematic prediction mistakes about failure to appear risk given defendant characteristics. Finally, I showed how this behavior analysis may inform the design of algorithmic decision systems by comparing expected social welfare under alternative algorithmic release rules against the observed release decisions of judges.

Empirical settings, such as pretrial release, medical diagnosis, and hiring, can serve as rich laboratories for behavioral analysis. My analysis provides a first step by exploring the empirical content of expected utility maximization, a canonical model of decision making under uncertainty, in these settings. An exciting direction is to explore the sharp, testable implications of alternative behavioral models such as rational inattention (e.g., Sims, 2003; Gabaix, 2014; Caplin and Dean, 2015) as well as forms of salience (e.g., Gennaioli and Shleifer, 2010; Bordalo et al., 2016; Bordalo, Gennaioli and Shleifer, 2021). Exploiting the full potential of these empirical settings is an important, policy-relevant agenda at the intersection of economic theory, applications of machine learning, and microeconometrics.

[35] See Frankel (2021) for a principal-agent analysis of delegating decisions to a misaligned decision maker who observes additional private information.

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.

Abaluck, Jason, Leila Agha, David C. Chan, Daniel Singer, and Diana Zhu. 2020. "Fixing Misallocation with Guidelines: Awareness vs. Adherence." NBER Working Paper No. 27467.

Albright, Alex. 2019. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions."

Allen, Roy, and John Rehbeck. 2020. "Satisficing, Aggregation, and Quasilinear Utility."

Andrews, Isaiah, Jonathan Roth, and Ariel Pakes. 2019. "Inference for Linear Conditional Moment Inequalities." NBER Working Paper No. 26374.

Apesteguia, Jose, and Miguel A. Ballester. 2015. "A Measure of Rationality and Welfare." *Journal of Political Economy*, 123(6): 1278–1310.

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. 2020. "Measuring Racial Discrimination in Algorithms." NBER Working Paper No. 28222.

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

Athey, Susan. 2017. "Beyond prediction: Using big data for policy problems." *Science*, 355(6324): 483–485.

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.

Babichenko, Yakov, Inbal Talgam-Cohen, Haifeng Xu, and Konstantin Zabarnyi. 2021. "Regret-Minimizing Bayesian Persuasion." arXiv preprint, arXiv:2105.13870.

Bai, Yuehao, Andres Santos, and Azeem M. Shaikh. 2021. "A Practical Method for Testing Many Moment Inequalities." *Journal of Business Economics and Statistics*.

Barocas, Solon, and Andrew Selbst. 2016. "Big data's disparate impact." *California Law Review*, 104: 671–732.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

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.

Benitez-Silva, Hugo, Moshe Buchinsky, and John Rust. 2004. "How Large are the Classification Errors in the Social Security Disability Award Process?" NBER Working Paper Series No. 10219.

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. 2019. "Counterfactuals with Latent Information."

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

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. 2020. "Inaccurate Statistical Discrimination: An Identification Problem." NBER Working Paper Series No. 25935.

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. 2019. "Artificial Intelligence and Behavioral Economics." In *The Economics of Artificial Intelligence: An Agenda*. , ed. Ajay Agrawal, Joshua Gans and Avi Goldfarb, 587–608. University of Chicago Press.

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

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review*, 106(5): 124–127.

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.

Chen, Irene Y., Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. 2020. "Probabilistic Machine Learning for Healthcare." arXiv preprint, arXiv:2009.11087.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2019. "Inference on Causal and Structural Parameters using Many Moment Inequalities." *The Review of Economic Studies*, 86(5): 1867–1900.

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.

Cowgill, Bo. 2018. "Bias and Productivity in Humans and Machines: Theory and Evidence."

Cox, Gregory, and Xiaoxia Shi. 2020. "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models."

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 Yang. 2019. "Proposals for Improving the U.S. Pretrial System." The Hamilton Project.

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

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.

Echenique, Federico, Kota Saito, and Taisuke Imai. 2021. "Approximate Expected Utility Rationalization." arXiv preprint, arXiv:2102.06331.

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.

Florios, Kostas, and Spyros Skouras. 2008. "Exact computation of max weighted score estimators." *Journal of Econometrics*, 146(1): 86–91.

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.

Gillis, Talia. 2019. "False Dreams of Algorithmic Fairness: The Case of Credit Pricing."

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.

Henry, Marc, Romuald Meango, and Ismael Mourifie. 2020. "Revealing Gender-Specific Costs of STEM in an Extended Roy Model of Major Choice."

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

Hull, Peter. 2021. "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making." NBER Working Paper Series No. 28503.

Imbens, Guido W. 2003. "Sensitivity to Exogeneity Assumptions in Program Evaluation." *American Economic Review*, 93(2): 126–132.

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.

Kamenica, Emir. 2019. "Bayesian Persuasion and Information Design." *Annual Review of Economics*, 11: 249–272.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review*, 101: 2590–2615.

Kitagawa, Toru, and Aleksey Tetenov. 2018. "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice." *Econometrica*, 86(2): 591–616.

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

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018*b*. "Algorithmic Fairness." *AEA Papers and Proceedings*, 108: 22–27.

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.

Kosterina, Svetlana. 2022. "Persuasion with unknown beliefs." *Theoretical Economics*, 17(3): 1075–1107.

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.

Lakkaraju, Himabindu, and Cynthia Rudin. 2017. "Learning Cost-Effective and Interpretable Treatment Regimes." *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54: 166–175.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." *KDD '17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

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.

Low, Hamish, and Luigi Pistaferri. 2015. "Disability Insurance and the Dynamics of the Incentive Insurance Trade-Off." *American Economic Review*, 105(10): 2986–3029.

Low, Hamish, and Luigi Pistaferri. 2019. "Disability Insurance: Error Rates and Gender Differences." NBER Working Paper No. 26513.

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. 1989. "Anatomy of the Selection Problem." *Journal of Human Resources*, 24(3): 343–360.

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. 2021. "A Robust Test of Prejudice for Discrimination Experiments."

Meehl, Paul E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* University of Minnesota Press.

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

Mourifie, Ismael, Marc Henry, and Romuald Meango. 2019. "Sharp bounds and testability of a Roy model of STEM major choices." arXiv preprint arXiv:1709.09284.

Mullainathan, Sendhi, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives*, 31(2): 87–106.

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, and Jonathan Roth. 2020. "An Honest Approach to Parallel Trends."

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

Rehbeck, John. 2020. "Revealed Bayesian Expected Utility with Limited Data."

Ribers, Michael Allan, and Hannes Ullrich. 2019. "Battling Antibiotic Resistance: Can Machine Learning Improve Prescribing?" arXiv preprint arXiv:1906.03044.

Rosenbaum, Paul R. 2002. *Observational Studies.* Springer.

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

Shapiro, Alexander. 1991. "Asymptotic Analysis of Stochastic Programs." *Annals of Operations Research*, 30: 169–186.

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. 2019. "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.

Tamer, Elie. 2003. "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria." *The Review of Economic Studies*, 70(1).

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

Viviano, Davide. 2020. "Policy Targeting under Network Interference."

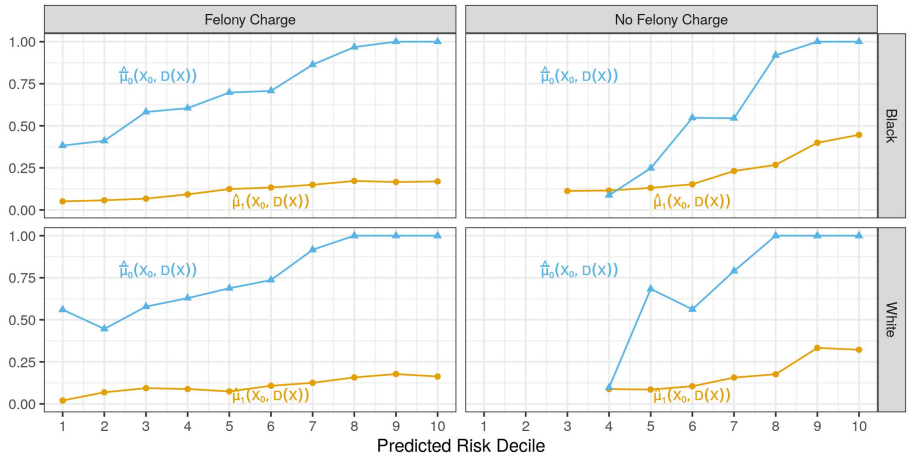
Wilder, Bryan, Eric Horvitz, and Ece Kamar. 2020. "Learning to Complement Humans." 1526–1533. International Joint Conferences on Artificial Intelligence Organization.

Wright, Marvin N., and Andreas Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software, Articles*, 77(1): 1–17.

Yang, Crystal, and Will Dobbie. 2020. "Equal Protection Under Algorithms: A New Statistical and Legal Framework." *Michigan Law Review*, 119(2): 291–396.

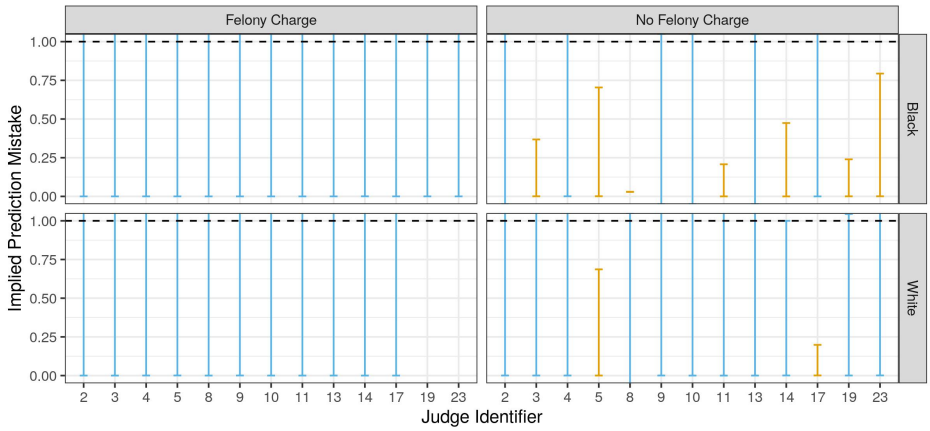
A Additional figures and tables

Figure A1: Observed failure to appear rate among released defendants and bound on the failure to appear rate among detained defendants by race-and-felony charge cells for one judge in New York City.



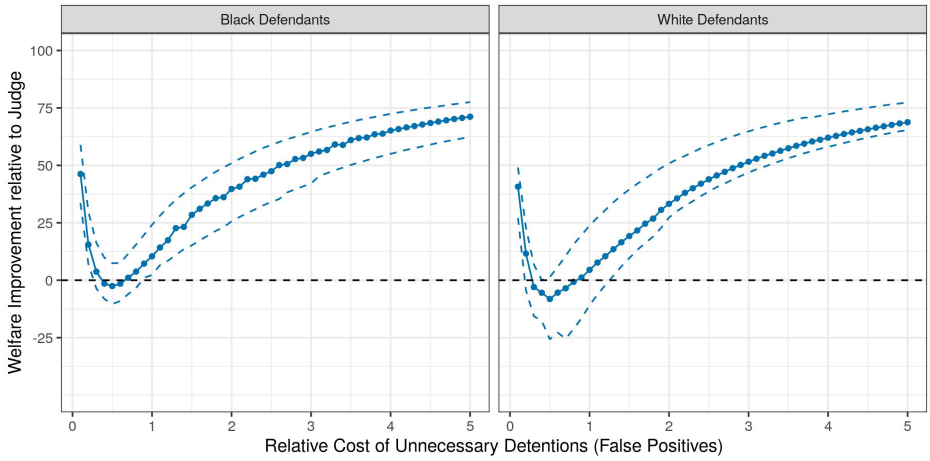
Notes: This figure plots the observed failure to appear rate among released defendants (orange, circles) and the bounds based on the judge leniency for the failure to appear rate among detained defendants (blue, triangles) at each decile of predicted failure to appear risk and race-by-felony charge cell 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 further estimation details on these bounds.

Figure A2: Estimated bounds on implied prediction mistakes between top and bottom predicted failure to appear risk deciles made by judges within each race-by-felony charge cell.



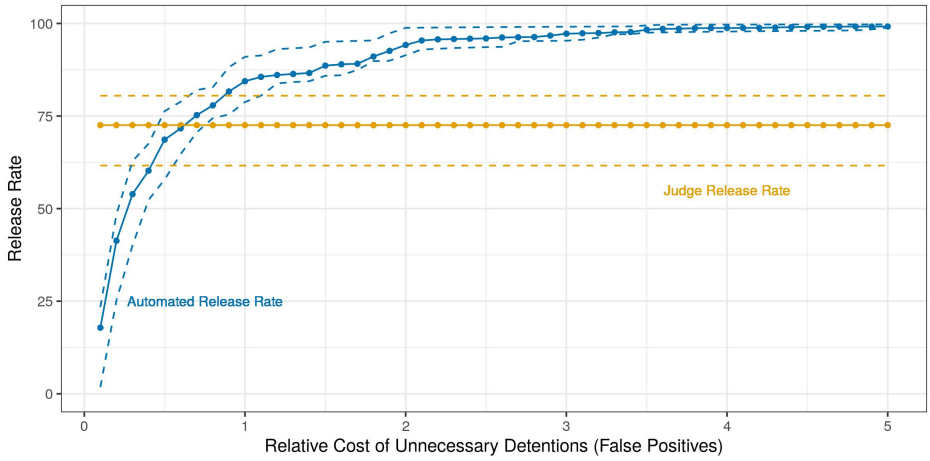
Notes: This figure plots the 95% confidence interval on the implied prediction mistake $\delta(x_0, d)/\delta(x_0, d')$ between the top decile d and bottom decile d' of the predicted failure to appear risk distribution for each judge in the top 25 whose pretrial release decisions violated the implied revealed preference inequalities (Table 1) 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 4.2.1 for theoretical details on the implied prediction mistake and Section 5.5 for the estimation details.

Figure A3: Comparison of full automation algorithmic decision rule relative to decisions of judges that made systematic prediction mistakes, separately by defendant race.



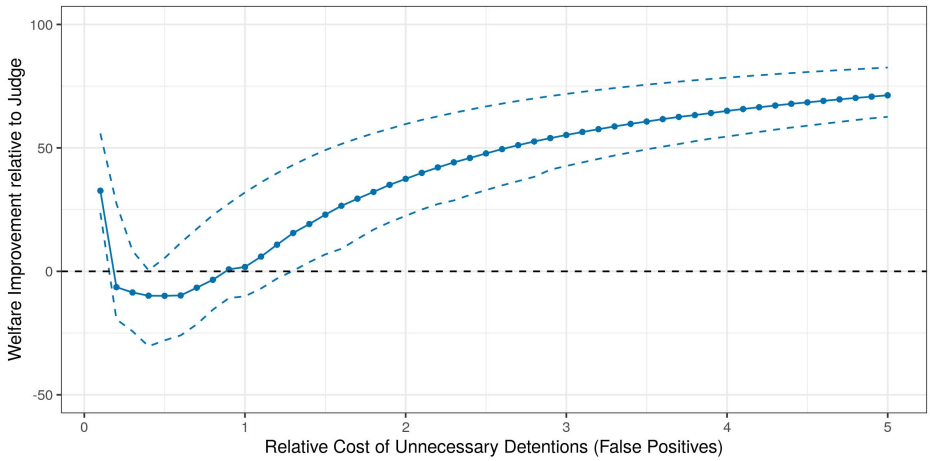
Notes: This figure reports the change in worst-case expected social welfare under the algorithmic decision rule that fully automates decisions against the observed release decisions of judges who made systematic prediction mistakes, separately by defendant race. The x-axis plots the relative social welfare cost of detaining a defendant that would not fail to appear in court $|\tilde{u}|$ (i.e., an unnecessary detention), where $u^*_{0,1} = -\tilde{u}/|1 + \tilde{u}|$. 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 further details.

Figure A4: Release rates under full automation algorithmic decision rule relative to the release rates of judges that made systematic prediction mistakes.



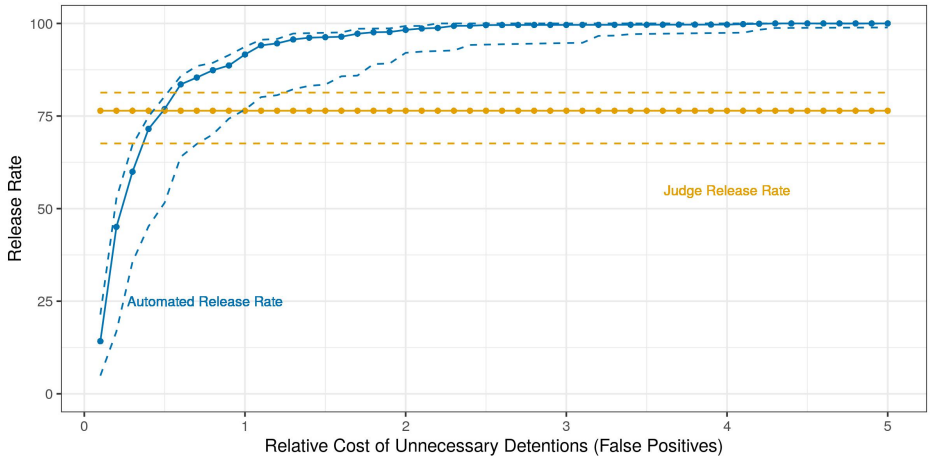
Notes: This figure reports the overall release rate of the algorithmic decision rule that fully automates decisions against the release rates of judges that made systematic prediction mistakes. These decisions rules are constructed and evaluated over race-by-age cells and deciles of predicted risk. The 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 $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$. 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 further details.

Figure A5: Comparison of full automation algorithmic decision rule relative to release decisions of judges that do not make systematic prediction mistakes.



Notes: This figure reports the change in worst-case expected social welfare under the algorithmic decision rule that fully automates decision-making against the observed release decisions of judges whose choices were consistent with expected utility maximization at accurate beliefs about 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, and the dashed lines report the minimum and maximum change across judges. See Section 6 for further details.

Figure A6: Release rates under algorithmic decision rule relative to the release rates of judges that did not make systematic prediction mistakes.



Notes: This figure reports the release rate of the algorithmic decision rule that fully automates decisions against the observed release rates among judges whose choices were consistent with expected utility maximization behavior at accurate beliefs. 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 $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$. 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 further details.

B Omitted proofs

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

B.1.1 Proof of Theorem 3.1

Proof. To prove this result, I first establish the following lemma.

Lemma B.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. $\mu_1(x_0, x_1) \leq \sum_{k=1}^K |u_{0,k}(x_0)|$ *for all* $(x_0, x_1) \in \mathcal{X}_0 \times \mathcal{X}_1$ *with* $\pi_1(x_0, x_1) > 0$,

ii. $\sum_{k=1}^K |u_{0,k}(x_0)| \leq \overline{\mu}_0(x_0, x_1)$ *for all* $(x_0, x_1) \in \mathcal{X}_0 \times \mathcal{X}_1$ *with* $\pi_0(x_0, x_1) > 0$.

Proof. This is an immediate consequence of applying Theorem C.1 to a screening decision with a binary choice. For all $(x_0, x_1) \in \mathcal{X}_0 \times \mathcal{X}_1$ with $\pi_1(x_0, x_1) > 0$, Theorem C.1 requires $\mu_1(x_0, x_1) \leq \sum_{k=1}^K |u_{0,k}(x_0)|$. For all $(x_0, x_1) \in \mathcal{X}_0 \times \mathcal{X}_1$ with $\pi_0(x_0, x_1) > 0$, Theorem C.1 requires $\sum_{k=1}^K |u_{0,k}(x_0)| \leq \mu_0(x_0, x_1)$. 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 B.1, the decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a linear utility function $u(c, y^*; x_0)$ satisfying, for all $x_0 \in \mathcal{X}_0$,

$$\max_{x_1 \in \Pi_1(x_0)} \mu_1(x_0, x_1) \leq \sum_{k=1}^K |u_{0,k}(x_0)| \leq \min_{x_1 \in \Pi_0(x_0)} \overline{\mu}_0(x_0, x_1)$$

The inequalities in Theorem 3.1 and the characterization of the identified set of linear utility functions in Corollary 3.1 are immediate. □

B.1.2 Proof of Proposition 3.1

Proof. 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 2, $\mu(x, z) = \mu(x, \tilde{z}) = \mu(x)$ for all $x \in \mathcal{X}$ and $z, \tilde{z} \in \mathcal{Z}$. Furthermore, using the fact that $Y_k^* \in [0, 1]$ for all $k = 1, \ldots, K$ and applying worst-case bounds (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 then 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. □

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

B.2.1 Proof of Theorem 4.1

Proof. This follows from applying Theorem C.2 to a screening decision with a binary choice over the class of linear utility functions. For all $(x_0, x_1) \in \mathcal{X}_0 \times \mathcal{X}_1$ with $\pi_1(x_0, x_1) > 0$, Theorem C.2 requires $\mu_1(x_0, x_1) - \epsilon(x_0, x_1) \leq \sum_{k=1}^{K} |u_{0,k}(x_0)|$. For all $(x_0, x_1) \in \mathcal{X}_0 \times \mathcal{X}_1$ with $\pi_0(x_0, x_1) > 0$, Theorem C.2 requires $\sum_{k=1}^{K} |u_{0,k}(x_0)| \leq \overline{\mu}_0(x_0, x_1) + \epsilon(x_0, x_1)$. 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 $x_0 \in \mathcal{X}_0$,

$$\max_{x_1 \in \mathcal{X}_1} \{\mu_1(x_0, x_1) - \epsilon(x_0, x_1)\} \leq \min_{x_1' \in \mathcal{X}_1} \{\overline{\mu}_0(x_0, x_1') + \epsilon(x_0, x_1')\}.$$

This is equivalent to, for all $x_0 \in \mathcal{X}_0$,

$$\mu_1(x_0, x_1) - \overline{\mu}_0(x_0, x_1') - \epsilon(x_0, x_1) - \epsilon(x_0, x_1') \leq 0 \text{ for all } x_1, x_1' \in \mathcal{X}_1.$$

□

B.2.2 Proof of Theorem 4.2

Proof. To prove this result, I will show that $1 - \underline{\lambda} = P(\overline{\mathcal{X}}_R)$. 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 $(x_0, x_1), (x_0, x_1') \in \mathcal{X}_R$, $\mu_1(x_0, x_1) - \overline{\mu}_0(x_0, x_1') \leq 0$ by construction. As a consequence, $P(\overline{\mathcal{X}}_R) \geq P(\mathcal{X}_R) = \sum_{x \in \mathcal{X}} P(x) 1\{\epsilon^*(x) = 0\} = 1 - \underline{\lambda}$.

Next, for each $x \in \overline{\mathcal{X}}_R$, define $\overline{\epsilon}(x) = 0$. For each $x = (x_0, x_1) \notin \overline{\mathcal{X}}_R$, define

$$\overline{\epsilon}_1(x) = \max_{x_1'} \{\mu_1(x) - \overline{\mu}_0(x_0, x_1')\}, \text{ and } \overline{\epsilon}_2(x) = \max_{x_1'} \{\mu_1(x_0, x_1') - \overline{\mu}_0(x)\},$$

and set $\overline{\epsilon}(x) = \max\{\overline{\epsilon}_1(x), \overline{\epsilon}_2(x)\}$. By construction, $\mu_1(x_0, x_1) - \overline{\mu}_0(x_0, x_1') - \overline{\epsilon}(x_0, x_1) - \overline{\epsilon}(x_0, x_1') \leq 0$ for all pairs $(x_0, x_1), (x_0, x_1') \in \mathcal{X}$. $\overline{\epsilon} = \{\overline{\epsilon}(x) \colon x \in \mathcal{X}\}$ 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 $1 - \underline{\lambda} \geq 1 - \sum_{x \in \mathcal{X}} P(x) 1\{\overline{\epsilon}(x) > 0\} = \sum_{x \in \mathcal{X}} P(x) 1\{\overline{\epsilon}(x) = 0\} = P(\overline{\mathcal{X}}_R)$. □

B.2.3 Proof of Theorem 4.3

Proof. This is an immediate consequence of applying Theorem C.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 C.3

requires

$$\mathbb{E}_{\widetilde{P}}[\omega(Y^*;X)\bar{Y}^* \mid C=1, X=x] \leq \mathbb{E}_{\widetilde{P}}[\omega(Y^*;X) \mid C=1, X=x] \left(\sum_{k=1}^{K} |u_{0,k}(x_0)|\right).$$

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

$$\mathbb{E}_{\widetilde{P}}[\omega_1(Y^*;X)\bar{Y}^* \mid C=1, X=x] \leq \sum_{k=1}^{K} |u_{0,k}(x_0)|.$$

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

$$\sum_{k=1}^{K} |u_{0,k}(x_0)| \leq \mathbb{E}_{\widetilde{P}}[\omega_0(Y^*;X)\bar{Y}^* \mid C=0, X=x],$$

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, $\widetilde{P}(\cdot \mid x) \in \mathcal{B}_{0,x}$ for all $x \in \mathcal{X}$ and non-negative weights $\omega(y^*;x)$ satisfying, for all $x_0 \in \mathcal{X}_0$,

$$\max_{\tilde{x}_1 \in \Pi_1(x_0)} \mathbb{E}_{\widetilde{P}}[\omega_1(Y^*;X)\bar{Y}^* \mid C=1, X=x] \leq \sum_{k=1}^{K} |u_{0,k}(x_0)|,$$

$$\sum_{k=1}^{K} |u_{0,k}(x_0)| \leq \min_{\tilde{x}_1 \in \Pi_0(x_0)} \mathbb{E}_{\widetilde{P}}[\omega_0(Y^*;X)\bar{Y}^* \mid C=0, X=x]$$

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

B.2.4 Proof of Theorem 4.4

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

$$\omega(1;x)u_{1,1}(x_0)P_1(1 \mid x) \geq \omega(0;x)u_{0,1}(x_0)P_1(0 \mid x),$$
$$\omega(0;x)u_{0,1}(x_0)\widetilde{P}_0(0 \mid x) \geq \omega(1;x)u_{1,1}(x_0)\widetilde{P}_0(1 \mid x),$$

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

$$P_1(1 \mid x) \leq \frac{\omega(0;x)u_{0,1}(x_0)}{\omega(0;x)u_{0,1}(x_0) + \omega(1;x)u_{1,1}(x_0)} \leq \widetilde{P}_0(1 \mid x).$$

The result follows by applying the bounds on $\widetilde{P}_0(1 \mid x)$ in a screening decision with a binary outcome. □

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

B.3.1 Proof of Proposition 5.1

Proof. This is an immediate consequence of applying Lemma C.1 to a binary choice, screening decision and iterated expectations. Since the decision maker's choices are consistent with expected utility maximization, by Lemma C.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 $x \in \mathcal{X}$,

$$\sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X = x) u(1, y^*; x_0) \geq \sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X = x) u(0, y^*; x_0)$$
$$\sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X = x) u(0, y^*; x_0) \geq \sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X = x) u(1, y^*; x_0),$$

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 $d \in \{1, \ldots, N_d\}$,

$$\sum_{x_1 \colon D(x_0, x_1) = d} \sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X = (x_0, x_1)) \frac{P(X_1 = x_1 \mid X_0 = x_0)}{P(D(X_0, X_1) = d \mid X_0 = x_0)} u(1, y^*; x_0) \geq$$

$$\sum_{x_1 \colon D(x_0, x_1) = d} \sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X = (x_0, x_1)) \frac{P(X_1 = x_1 \mid X_0 = x_0)}{P(D(X_0, X_1) = d \mid X_0 = x_0)} u(0, y^*; x_0)$$

and

$$\sum_{x_1 \colon D(x_0, x_1) = d} \sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X = (x_0, x_1)) \frac{P(X_1 = x_1 \mid X_0 = x_0)}{P(D(X_0, X_1) = d \mid X_0 = x_0)} u(0, y^*; x_0) \geq$$

$$\sum_{x_1 \colon D(x_0, x_1) = d} \sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X = (x_0, x_1)) \frac{P(X_1 = x_1 \mid X_0 = x_0)}{P(D(X_0, X_1) = d \mid X_0 = x_0)} u(1, y^*; x_0).$$

These can equivalently be written as

$$\sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X_0 = x_0, D(X) = d) u(1, y^*; x_0) \geq$$

$$\sum_{y^* \in \mathcal{Y}} P(Y^* = y^*, C = 1 \mid X_0 = x_0, D(X) = d) u(0, y^*; x_0)$$

and

$$\sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X_0 = x_0, D(X) = d) u(0, y^*; x_0) \geq$$

$$\sum_{y^* \in \mathcal{Y}} \widetilde{P}(Y^* = y^*, C = 0 \mid X_0 = x_0, D(X) = d) u(1, y^*; x_0).$$

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

B.3.2 Proof of Proposition 5.2

Proof. Under the behavioral restriction $V \mid \{Y^*, X\} \sim V \mid \{Y^*, X_0, \mu(X)\}$ and $C \mid \{V, X\} \sim C \mid \{V, X_0, \mu(X)\}$, 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 $\tilde{Q}(X)\tilde{Q}(Y^* \mid X)\tilde{Q}(V \mid Y^*, \mu(X), X_0)\tilde{Q}(C \mid V, \mu(X), X_0)$. The result then follows by the same argument as the proof of Theorem 3.1. □

C 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 of these assumptions, and analyze *treatment assignment problems* in which the decision maker selects one of many treatments for each individual. This nests the main text characterization of expected utility maximization in screening decisions as a special case.

C.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* $y_j := y(c_j) = (y_1(c_j), \ldots, y_K(c_j)) \in \mathcal{Y} \subseteq [0,1]^K$ 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 $(W, X, C, \vec{Y}) \sim P(\cdot)$ 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 $\mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_x)$, where $\mathcal{B}_x := \{\mathcal{B}_{j,x} \colon j = 1, \ldots, J\}$.

Definition 7. The *utility function* $u \colon \{c_1, \ldots, c_J\} \times \mathcal{Y}^J \times \mathcal{X}_0$ specifies the payoff associated with each choice, vector of potential outcomes, and characteristics $x_0 \in \mathcal{X}_0$. Let $\mathcal{U}$ denote the feasible set of utility functions specified by the researcher.

Definition 8. The decision maker's choices are *consistent with expected utility maximization* in a treatment assignment problem if there exists $u \in \mathcal{U}$ and $(X, V, C, \vec{Y}) \sim Q$ 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 $Q(c_j \mid x, v) > 0$,

$$\mathbb{E}_Q \left[u(c_j, \vec{Y}; X_0) \mid X = x, V = v \right] \geq \mathbb{E}_Q \left[u(c', \vec{Y}; X_0) \mid X = x, V = v \right]$$

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

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

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*, denoted $\mathcal{H}_P(u; \mathcal{B}) \subseteq \mathcal{U}$, is the set of $u \in \mathcal{U}$ such that there exists $(X, V, C, \vec{Y}) \sim Q$ satisfying (i)-(iii).

C.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 C.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*

$$\mathbb{E}_Q \left[u(c_j, \vec{Y}; X_0) \mid C = c_j, X = x \right] \geq \mathbb{E}_Q \left[u(c', \vec{Y}; X_0) \mid C = c_j, X = x \right] \tag{18}$$

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

Corollary C.1. *The identified set of utility functions $\mathcal{H}_P(u; \mathcal{B})$ 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 (18).*

Theorem C.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 discussed in the main text, the key insight in proving Theorem C.1 is that checking whether behavior is equivalent with EU maximization is equivalent to an information design problem (Bergemann and Morris, 2019; Kamenica, 2019). I must simultaneously check whether the information designer could induce the observed choices at *any* accurate beliefs in the identified set due to the missing data problem, and *any* utility function in the researcher-specified feasible set of utility function $\mathcal{U}$ since the decision maker's true payoffs are unknown.

As in Section 3, I next analyze the testable implications of expected utility maximization behavior for a binary choice $c \in \{0, 1\}$ over linear utility functions of the form $u(c, \vec{y}; x_0) = \sum_{k=1}^{K} y_k - u_{0,k}(x_0)c$, where $u_{0,k}(x_0) \geq 0$ for all $x_0 \in \mathcal{X}_0$.[36] As in the main text, define $\Pi_1(x_0) :=$

[36] Over the class of linear utility functions, the expected utility maximization model in a treatment assignment

$\{x_1 \colon \pi_1(x_0, x_1) > 0\}$, $\Pi_0(x_0) := \{x_1 \colon \pi_0(x_0, x_1) > 0\}$. 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 C.2. *The decision maker's choices are consistent with expected utility maximization at some linear utility function if and only if, for all $x_0 \in \mathcal{X}_0$,*

$$\max_{x_1 \in \Pi_0(x_0)} \underline{\mu}_0(x_0, x_1) \leq \min_{x_1 \in \Pi_1(x_0)} \overline{\mu}_1(x_0, x_1). \tag{19}$$

The identified set of linear utility functions $\mathcal{H}_P(u; \mathcal{B})$ equals the set of all utility functions satisfying, for all $x_0 \in \mathcal{X}_0$, $u(c, \vec{y}; x_0) = \sum_{k=1}^{K} y_k - u_{0,k}(x_0)c$ with $u_{0,k}(x_0) \geq 0$ and

$$\max_{x_1 \in \Pi_0(x_0)} \underline{\mu}_0(x_0, x_1) \leq \sum_{k=1}^{K} |u_{0,k}(x_0)| \leq \min_{x_1 \in \Pi_1(x_0)} \overline{\mu}_1(x_0, x_1). \tag{20}$$

Corollary C.2 immediately implies two negative results about the identification of systematic prediction mistakes in treatment assignment problems that parallel Corollary 3.2 in the main text for screening decisions.

Corollary C.3. *The decision maker's choices are consistent with expected utility maximization behavior at some linear utility function $u(c, \vec{y}; x_0) = \sum_{k=1}^{K} y_k + u_{0,k}(x_0)c$ if either:*

(i) All characteristics affect utility (i.e., $\mathcal{X} = \mathcal{X}_0$) and $\underline{\mu}_0(x_0) \leq \overline{\mu}_1(x_0)$ for all $x_0 \in \mathcal{X}_0$.

(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 $y_c \in \mathcal{Y}$.

C.3 Approximate expected utility maximization in treatment assignment problems

I now 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 5) again generalizes naturally to treatment assignment problems.

Definition 9. 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 $(X, V, C, \vec{Y}) \sim Q$ 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 $Q(c \mid x, v) > 0$,

$$\mathbb{E}_Q\left[u(c, \vec{Y}; X_0) \mid X = x, V = v\right] \geq \mathbb{E}_Q\left[u(c', \vec{Y}; X_0) \mid X = x, V = v\right] - \epsilon(x),$$

problem is an extended Roy model in which the expected benefit function only depends on the realized outcome and the cost function only depends on the choice. Henry, Meango and Mourifie (2020) also studies extended Roy model behavior under the additional assumption that the utility function satisfies $u(0, \vec{y}; x_0) = y_0, u(1, \vec{y}; x_0) = Y_1 - \lambda(Y_1)$ for some function $\lambda(\cdot)$.

and (ii) Information Set, (iii) Data Consistency as defined in Definition 8. The *identified set of expected utility costs*, denoted $\mathcal{H}_P(\epsilon;\mathcal{B}) \subseteq \mathcal{U}$, is the set of $\epsilon = \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ such that there exists $u \in \mathcal{U}$ and $(X, V, C, \vec{Y}) \sim Q$ satisfying (i)-(iii).

I provide a necessary and sufficient characterization of approximate expected utility maximization at accurate beliefs in treatment assignment problems.

Theorem C.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*

$$\mathbb{E}_Q\left[u(c_j, \vec{Y}; X_0) \mid C = c, X = x\right] \geq \mathbb{E}_Q\left[u(c', \vec{Y}; X_0) \mid C = c, X = x\right] - \epsilon(x) \tag{21}$$

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

Corollary C.4. *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 $x = (x_0, x_1)$, $x' = (x_0, x_1')$,*

$$\underline{\mu}_0(x) - \overline{\mu}_1(x') - \epsilon(x) - \epsilon(x') \leq 0. \tag{22}$$

$\mathcal{H}_P(\epsilon;\mathcal{B})$ equals the set of all $\epsilon = \{\epsilon(x) \geq 0 \colon x \in \mathcal{X}\}$ satisfying (22).

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

I now characterize conditions under which the decision maker's choices are consistent with expected utility maximization at inaccurate beliefs in a treatment assignment problem. The definition of expected utility maximization behavior at inaccurate beliefs in the main text (Definition 6) again generalizes naturally to treatment assignment problems by modifying the Data Consistency condition.

Definition 10. 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 $(X, V, C, \vec{Y}) \sim Q$ satisfying (i) Expected Utility Maximization, (ii) Information Set as in Definition 8, 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 C.3. *Assume $\widetilde{P}(\cdot \mid x) > 0$ for all $\widetilde{P}(\cdot \mid x) \in \mathcal{H}_P(P(\cdot \mid x);\mathcal{B}_x)$ 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 $\pi_j(x) > 0$, $c' \neq c_j$

$$\mathbb{E}_{\widetilde{P}}\left[\omega(\vec{Y}; X)u(c_j, \vec{Y}; X_0) \mid C = c_j, X = x\right] \geq \mathbb{E}_{\widetilde{P}}\left[\omega(\vec{Y}; X)u(c', \vec{Y}; X_0) \mid C = c_j, X = x\right]$$

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 $(X, C, \vec{Y}) \sim \widetilde{P}$ defined as $\widetilde{P}(x, c_j, \vec{y}) = \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)P(x)$.

C.5 Proofs of characterization results for treatment assignment problems

C.5.1 Proof of Theorem C.1

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

Lemma C.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*

$$\sum_{\vec{y} \in \mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)u(c_j, \vec{y}; x_0) \geq \sum_{\vec{y} \in \mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid x)\pi_j(x)u(c', \vec{y}; x_0)$$

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

Proof of Lemma C.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 C.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 $(x, v) \in \mathcal{X} \times \mathcal{V}$,

$$
Q(c_j \mid x, v) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c_j, \vec{y}; x_0) \right) \geq Q(c_j \mid x, v) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c', \vec{y}; x_0) \right)
$$

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

$$
Q(c_j \mid x, v) Q(v \mid x) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c_j, \vec{y}; x_0) \right) \geq Q(c_j \mid x, v) Q(v \mid x) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c', \vec{y}; x_0) \right).
$$

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

$$
Q(v \mid x) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x, v) u(c_j, \vec{y}; x_0) \right) \geq Q(v \mid x) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j, \vec{y} \mid x, v) u(c', \vec{y}; x_0) \right).
$$

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

$$\sum_{\vec{y}\in\mathcal{Y}^J}\left(\sum_{v\in\mathcal{V}}Q(v,c_j,\vec{y}\mid x)\right)u(c_j,\vec{y};x_0)\geq\sum_{\vec{y}\in\mathcal{Y}^J}\left(\sum_{v\in\mathcal{V}}Q(v,c_j,\vec{y}\mid x)\right)u(c',\vec{y};x_0)$$
$$\sum_{\vec{y}\in\mathcal{Y}^J}Q(c_j,\vec{y}\mid x)u(c_j,\vec{y};x_0)\geq\sum_{\vec{y}\in\mathcal{Y}^J}Q(c_j,\vec{y}\mid x)u(c',\vec{y};w).$$

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

Proof of Lemma C.1: Sufficiency As notation, let $\mathcal{C}:=\{c_1,\ldots,c_J\}$. To establish sufficiency, I show that if the conditions in Lemma C.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 C.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 $v\in\mathcal{V}:=\left\{1,\ldots,2^J\right\}$ 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

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

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

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

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 C.1 were assumed to hold and data consistency holds.

Lemma C.1 implies Theorem C.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 C.1 as: for all $c_j \in \{c_1, \ldots, c_J\}$ and $c' \neq c_j$,

$$\sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x)u(c_j, \vec{y}; x_0) \geq \sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x_0)u(c', \vec{y}; x_0).$$

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

$$\sum_{\vec{y}\in\mathcal{Y}^J} Q_j(\vec{y} \mid x)u(c_j, \vec{y}; x_0) \geq \sum_{\vec{y}\in\mathcal{Y}^J} Q_j(\vec{y} \mid x_0)U(c', \vec{y}; w).$$

The statement of Theorem C.1 follows by noticing that

$$\sum_{\vec{y}\in\mathcal{Y}^J} Q_j(\vec{y} \mid x)u(c_j, \vec{y}; x_0) = \mathbb{E}_Q\left[U(c_j, \vec{Y}; x_0) \mid C = c_j, X = x\right],$$
$$\sum_{\vec{y}\in\mathcal{Y}^J} Q_j(\vec{y} \mid x)U(c', \vec{y}; x_0) = \mathbb{E}_Q\left[U(c', \vec{Y}; x_0) \mid C = c_j, X = x\right].$$

□

C.5.2 Proof of Corollary C.2

Proof. This follows from Theorem C.1. For all $x \in \mathcal{X}$ with $\pi_1(x) > 0$, the stochastic revealed preference inequalities require $\sum_{k=1}^{K} \mathbb{E}[Y_k(1) \mid C = 1, X = x] - u_{0,k}(x_0) \geq \sum_{k=1}^{K} \mathbb{E}[Y_k(0) \mid C = 1, X = x]$. For all $x \in \mathcal{X}$ with $\pi_0(x) > 0$, the stochastic revealed preference inequalities require $\sum_{k=1}^{K} \mathbb{E}[Y_k(0) \mid C = 0, X = x] \geq \sum_{k=1}^{K} \mathbb{E}[Y_k(1) \mid C = 0, X = x] - u_{0,k}(x_0)$. 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

$$\mu_0(x_0, x_1) \leq \sum_{k=1}^{K} |u_{0,k}(x_0)| \text{ whenever } \pi_0(x) > 0$$

$$\sum_{k=1}^{K} |u_{0,k}(x_0)| \leq \mu_1(x_0, x_1) \text{ whenever } \pi_1(x) > 0.$$

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. □

C.5.3 Proof of Theorem C.2

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

Lemma C.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*

$$\sum_{\vec{y} \in \mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid x) \pi_j(x) u(c_j, \vec{y}; x_0) \geq \sum_{\vec{y} \in \mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid x) \pi_j(x) u(c', \vec{y}; x_0) - \pi_j(x) \epsilon(x)$$

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

Proof of Lemma C.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 C.1, it is without loss of generality to assume $V \in \mathcal{V}$ has finite support.

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

$$Q(c_j \mid x, v) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c_j, \vec{y}; x_0) \right) \geq Q(c_j \mid x, v) \left(\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid x, v) u(c', \vec{y}; x_0) \right) - Q(c_j \mid x, v) \epsilon(x)$$

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

$$Q(c_j \mid x, v)Q(v \mid x)\left(\sum_{\vec{y}\in\mathcal{Y}^J} Q(\vec{y} \mid x, v)u(c_j, \vec{y}; x_0)\right) \geq$$

$$Q(c_j \mid x, v)Q(v \mid x)\left(\sum_{\vec{y}\in\mathcal{Y}^J} Q(\vec{y} \mid x, v)u(c', \vec{y}; x_0)\right) - Q(c_j \mid x, v)Q(v \mid x)\epsilon(x).$$

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

$$Q(v \mid x)\left(\sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x, v)u(c_j, \vec{y}; x_0)\right) \geq Q(v \mid x)\left(\sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x, v)u(c', \vec{y}; x_0)\right) - Q(c_j, v \mid x)\epsilon(x)$$

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

$$\sum_{\vec{y}\in\mathcal{Y}^J}\left(\sum_{v\in\mathcal{V}} Q(v, c_j, \vec{y} \mid x)\right)u(c_j, \vec{y}; x_0) \geq \sum_{\vec{y}\in\mathcal{Y}^J}\left(\sum_{v\in\mathcal{V}} Q(v, c_j, \vec{y} \mid x)\right)u(c', \vec{y}; x_0) - \sum_{v\in\mathcal{V}} Q(c_j, v \mid x)\epsilon(x),$$

$$\sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x)u(c_j, \vec{y}; x_0) \geq \sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x)u(c', \vec{y}; x_0) - Q(c_j \mid x)\epsilon(x)$$

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

Proof of Lemma C.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 C.1.

Lemma C.2 implies Theorem C.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 C.2 as: for all $c_j \in \{c_1, \ldots, c_J\}$ and $c' \neq c_j$,

$$\sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x)u(c_j, \vec{y}; x_0) \geq \sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x_0)u(c', \vec{y}; w) - Q(c_j \mid x)\epsilon(x).$$

Notice that if $\pi_j(x) = 0$, then $Q(c_j, \vec{y} \mid w, x) = 0$ 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

$$\sum_{\vec{y}\in\mathcal{Y}^J} Q_j(\vec{y} \mid x)u(c_j, \vec{y}; x_0) \geq \sum_{\vec{y}\in\mathcal{Y}^J} Q_j(\vec{y} \mid x_0)u(c', \vec{y}; x_0) - \epsilon(x).$$

The statement of Theorem C.2 follows by noticing that

$$\sum_{\vec{y}\in\mathcal{Y}^J} Q_j(\vec{y} \mid x)u(c_j, \vec{y}; x_0) = \mathbb{E}_Q\left[u(c_j, \vec{Y}; x_0) \mid C = c_j, X = x\right],$$
$$\sum_{\vec{y}\in\mathcal{Y}^J} Q_j(\vec{y} \mid x)u(c', \vec{y}; x_0) = \mathbb{E}_Q\left[u(c', \vec{Y}; x_0) \mid C = c_j, X = x\right].$$

□

C.5.4 Proof of Corollary C.4

Proof. This follows from applying Theorem C.2 to a binary choice, treatment assignment problem over the class of linear utility functions. For all $(x_0, x_1) \in \mathcal{X}_0 \times \mathcal{X}_1$ with $\pi_0(x_0, x_1) > 0$, Theorem C.2 requires $\underline{\mu}_0(x_0, x_1) - \epsilon(x_0, x_1) \leq \sum_{k=1}^K |u_{0,k}(x_0)|$. For all $(x_0, x_1) \in \mathcal{X}_0 \times \mathcal{X}_1$ with $\pi_1(x_0, x_1) > 0$, Theorem C.2 requires $\sum_{k=1}^K |u_{0,k}(x_0)| \leq \overline{\mu}_1(x_0, x_1) + \epsilon(x_0, x_1)$. 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 $x_0 \in \mathcal{X}_0$,

$$\max_{x_1\in\mathcal{X}_1} \left\{\underline{\mu}_0(x_0, x_1) - \epsilon(x_0, x_1)\right\} \leq \min_{x_1'\in\mathcal{X}_1} \left\{\overline{\mu}_1(x_0, x_1) + \epsilon(x_0, x_1)\right\}.$$

This is equivalent to, for all $x_0 \in \mathcal{X}_0$,

$$\underline{\mu}_0(x_0, x_1) - \overline{\mu}_1(x_0, x_1') - \epsilon(x_0, x_1) - \epsilon(x_0, x_1') \leq 0 \text{ for all } x_1, x_1' \in \mathcal{X}_1.$$

□

C.5.5 Proof of Theorem C.3

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

Lemma C.3. *Assume $\widetilde{P}(\cdot \mid x) > 0$ for all $\widetilde{P}(\cdot \mid x) \in \mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_x)$ 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 $c' \neq c_j$,*

$$\sum_{\vec{y}\in\mathcal{Y}^J} Q(\vec{y} \mid x)\widetilde{P}(c_j \mid \vec{y}, x)u(c_j, \vec{y}; x_0) \geq \sum_{\vec{y}\in\mathcal{Y}^J} Q(\vec{y} \mid x)\widetilde{P}(c_j \mid \vec{y}, x)u(c', \vec{y}; x_0),$$

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

Proof of Lemma C.3: Necessity First, by an analogous argument as given in the proof of necessity for Lemma C.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 C.1, I arrive at

$$\sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x)u(c, \vec{y}; x_0) \geq \sum_{\vec{y}\in\mathcal{Y}^J} Q(c_j, \vec{y} \mid x)u(c', \vec{y}; x_0).$$

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 C.3: Sufficiency To show sufficiency, suppose that the conditions in Lemma C.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 $\widetilde{P}(c \mid \vec{y}, w, x)$ 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 C.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

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

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 10). 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 C.3, we have that

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid v, x)u(c_j, \vec{y}; x_0) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q(\vec{y} \mid v, x)u(c', \vec{y}; x_0),$$

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

$$\sum_{\vec{y} \in \mathcal{Y}^J} \frac{Q(\vec{y} \mid x)}{\widetilde{P}(\vec{y} \mid x)}\widetilde{P}(c, \vec{y} \mid x)u(c, \vec{y}; x_0) \geq \sum_{\vec{y} \in \mathcal{Y}^J} \frac{Q(\vec{y} \mid x)}{\widetilde{P}(\vec{y} \mid x)}\widetilde{P}(c, \vec{y} \mid x)U(c', \vec{y}; x_0).$$

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

$$\sum_{\vec{y} \in \mathcal{Y}^J} \frac{Q(\vec{y} \mid x)}{\widetilde{P}(\vec{y} \mid x)}\widetilde{P}_j(\vec{y} \mid x)u(c, \vec{y}; x_0) \geq \sum_{\vec{y} \in \mathcal{Y}^J} \frac{Q(\vec{y} \mid x)}{\widetilde{P}(\vec{y} \mid x)}\widetilde{P}_j(\vec{y} \mid x)u(c', \vec{y}; x_0).$$

The result follows by noticing $\sum_{\vec{y}\in\mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid x)\frac{Q(\vec{y}|x)}{\widetilde{P}(\vec{y}|x)}u(c,\vec{y};x_0) = \mathbb{E}_{\widetilde{P}}\left[\frac{Q(\vec{y}|x)}{\widetilde{P}(\vec{y}|x)}u(c,\vec{y};x_0)\right]$ and defining the weights as $\omega(\vec{y};x) = \frac{Q(\vec{y}|x)}{\widetilde{P}(\vec{y}|x)}$. □

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 are suitably modified to account for the dimension reduction discussed in Section 5.2. As in the main text, let $D\colon \mathcal{X} \to \{1,\dots,N_d\}$ be a function that partitions the observable characteristics X into level sets $\{x \in \mathcal{X}\colon D(x) = d\}$.

D.1 Approximate expected utility maximization after dimension reduction

The identification result for expected utility maximization behavior with inaccurate beliefs extends to coarsening the excluded characteristics. First, for a treatment assignment problem, Theorem C.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. This follows from Lemma C.2 and the same iterated expectations argument as in the proof of Proposition C.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 $x_0 \in \mathcal{X}_0$, $d \in \{1,\dots,N_d\}$, $c_j \in \{c_1,\dots,c_J\}$ and $c' \neq c_j$,*

$$\sum_{\vec{y}\in\mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid (x_0,d))u(c_j,\vec{y};x_0) \geq \sum_{\vec{y}\in\mathcal{Y}^J} \widetilde{P}_j(\vec{y} \mid (x_0,d))u(c',\vec{y};x_0) - \bar{\epsilon}(x_0,d),$$

where

$$\begin{aligned}
\widetilde{P}(c_j,\vec{y} \mid (x_0,d)) &= \sum_{x_1\colon D(x_0,x_1)=d} \widetilde{P}(c_j,\vec{y} \mid (x_0,x_1))\frac{P(X_1 = x_1 \mid X_0 = x_0)}{P(D(X_0,X_1) = d \mid X_0 = x_0)},\\
\pi_j(x_0,d) &= \sum_{x_1\colon D(x_0,x_1)=d} \pi_j(x_0,x_1)\frac{P(X_1 = x_1 \mid X_0 = x_0)}{P(D(X_0,X_1) = d \mid X_0 = x_0)},\\
\widetilde{P}_j(\vec{y} \mid (x_0,d)) &= \widetilde{P}(c_j,\vec{y} \mid (x_0,d))/\pi_j(x_0,d),\\
\bar{\epsilon}(x_0,d) &= \pi_j(x_0,d)^{-1} \sum_{x_1\colon D(x_0,x_1)=d} \epsilon(x_0,x_1)\pi_j(x_0,x_1)\frac{P(X_1 = x_1 \mid X_0 = x_0)}{P(D(X_0,X_1) = d \mid X_0 = x_0)}.
\end{aligned}$$

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 $\bar{\epsilon}(x_0,d) \geq 0$ such that, for all pairs (x_0,d), (x_0,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 $\underline{\mathcal{E}}^*(D) = \sum_{x_0,d} P(x_0,d)\epsilon^*(x_0,d)$, where $\epsilon^*(x)$ is an optimal solution to the linear program defined in (9) of the main text, and $\epsilon^*(x_0,d) = \max_{x_1:\, D(x_0,x_1)=d} \epsilon^*(x_0,x_1)$. 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

$$\begin{aligned}\underline{\mathcal{E}}(D) := \min_{\epsilon(x_0,d)} & \sum_{x_0,d} P(x_0,d)\epsilon(x_0,D(x)) \\ \text{s.t. } & \epsilon(x_0,d) \geq 0 \text{ for all } x \in \mathcal{X}, \\ & \mu_1(x_0,d) - \overline{\mu}_0(x_0,d') - \epsilon(x_0,d) - \epsilon(x_0,d') \leq 0 \text{ for all pairs } (x_0,d),(x_0,d').\end{aligned}$$

It is immediate that $\underline{\mathcal{E}}(D) \leq \underline{\mathcal{E}}^*(D)$ since $\epsilon^*(x_0,d)$ is feasible in the program, and so $\underline{\mathcal{E}}(D)$ provides 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 $\underline{\lambda}^*(D) = \sum_{x_0,d} P(x_0,d)1\{\epsilon^*(x_0,d) > 0\}$ 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

$$\begin{aligned}\underline{\lambda}(D) := \min_{\epsilon(x_0,d)} & \sum_{x_0,d} P(x_0,d)1\{\epsilon(x_0,D(x)) > 0\} \\ \text{s.t. } & \epsilon(x_0,d) \geq 0 \text{ for all } x \in \mathcal{X}, \\ & \mu_1(x_0,d) - \overline{\mu}_0(x_0,d') - \epsilon(x_0,d) - \epsilon(x_0,d') \leq 0 \text{ for all pairs } (x_0,d),(x_0,d').\end{aligned}$$

Since $\epsilon^*(x_0,d)$ is feasible, it follows that $\underline{\lambda}(D) \leq \underline{\lambda}^*(D)$, and so, $\underline{\lambda}(D)$ provides 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

The identification result for expected utility maximization behavior with inaccurate beliefs extends to coarsening the excluded characteristics. I first show that, for a treatment assignment problem, Theorem C.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. This follows directly from Lemma C.3 and the same iterated expectations argument as in the proof of Proposition 5.1.

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 $x_0 \in \mathcal{X}_0$, $d \in \{1,\ldots,N_d\}$, $c_j \in \{c_1,\ldots,c_J\}$ and $c' \neq c_j$,*

$$\sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j,\vec{y} \mid x_0,d)u(c,\vec{y};x_0) \geq \sum_{\vec{y} \in \mathcal{Y}^J} Q(c_j,\vec{y} \mid x_0,d)u(c',\vec{y};x_0),$$

where

$$Q(c, \vec{y} \mid x_0, d) = \left(\sum_{x_1 :\, D(x_0,x_1)=d} \widetilde{P}(c \mid \vec{y}, (x_0, x_1)) Q(\vec{y} \mid (x_0, x_1)) P(x_1 \mid x_0) \right) / P(D(X_0, X_1) = d \mid X_0 = x_0),$$

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

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 $\omega(c, \vec{y}; x_0, d) \geq 0$ satisfying, for all $c_j \in \{c_1, \ldots, c_J\}$ with $c' \neq c_j$ and $x \in \mathcal{X}$,

$$\sum_{\vec{y} \in \mathcal{Y}^J} \omega(c_j, \vec{y}; x_0, d) \widetilde{P}(c_j, \vec{y} \mid x_0, D(x) = d) u(c_j, \vec{y}; x_0) \geq$$

$$\sum_{\vec{y} \in \mathcal{Y}^J} \omega(c_j, \vec{y}; x_0, d) \widetilde{P}(c_j, \vec{y} \mid x_0, D(x) = d) u(c', \vec{y}; x_0)$$

and $\mathbb{E}_{\widetilde{P}} \left[\omega(C, \vec{Y}; X_0, D(X)) \mid X_0 = x_0, D(X) = d \right] = 1$.

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 Theorem 4.4, this result may be applied to derive bounds on the decision maker's reweighed utility threshold through

$$P_1(1 \mid x_0, d) \leq \frac{\omega(0, 0; x_0, d) u_{0,1}(x_0)}{\omega(0, 0; x_0, d) u_{0,1}(x_0) + \omega(1, 1; x_0, d) u_{1,1}(x_0)} \leq \overline{P}_0(1 \mid x_0, d), \tag{23}$$

where $P_c(y^* \mid x_0, d) := P(Y^* = y^* \mid C = c, X_0 = x_0, D(X) = d)$. Next, define $M = 1\{C = 0, Y^* = 0\} + 1\{C = 1, Y^* = 1\}$, $\tau(x_0, d) = \frac{\omega(0,0;x_0,d)u_{0,1}(x_0)}{\omega(0,0;x_0,d)u_{0,1}(x_0) + \omega(1,1;x_0,d)u_{1,1}(x_0)}$. Examining $x_0 \in \mathcal{X}_0$, $d, d' \in \{1, \ldots, N_d\}$, we arrive at

$$\frac{(1 - \tau(x_0, d))/\tau(x_0, d)}{(1 - \tau(x_0, d'))/\tau(x_0, d')} = \frac{\frac{Q(C=1,Y^*=1|M=1,x_0,d)/Q(C=0,Y^*=0|M=1,x_0,d)}{Q(C=1,Y^*=1|M=1,x_0,d')/Q(C=0,Y^*=0|M=1,x_0,d')}}{\frac{P(C=1,Y^*=1|M=1,x_0,d)/P(C=0,Y^*=0|M=1,x_0,d)}{P(C=1,Y^*=1|M=1,x_0,d')/P(C=0,Y^*=0|M=1,x_0,d')}}. \tag{24}$$

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_0, d) and (x_0, d'). If these bounds lie everywhere above one, then the decision maker's beliefs about their own ex-post errors are overreacting.

E Additional results for the econometric framework

E.1 Quasi-randomly assigned instrumental variable

I modify Assumption 2 to only impose that the instrument be quasi-randomly assigned conditional on some additional characteristics $t \in \mathcal{T}$ with finite support. The joint distribution $(X, T, Z, C, Y^*) \sim P$ satisfies

$$(X, Y^*) \perp\!\!\!\perp Z \mid T \tag{25}$$

and $P(x, t, z) > 0$ for all $(x, t, z) \in \mathcal{X} \times \mathcal{T} \times \mathcal{Z}$. In the empirical application to the New York City pretrial system, judges are quasi-randomly assigned to cases within a court-by-time cell.

Under (25), researchers can derive bounds on the unobservable conditional outcome probabilities. Let $\mu(x, z) := \mathbb{E}[Y^* \mid X = x, Z = z]$ and $\mu(x, z, t) := \mathbb{E}[Y^* \mid X = x, Z = z, T = t]$. Then, by iterated expectations,

$$\mu(x, z) = \sum_{t \in \mathcal{T}} \mu(x, z, t) P(t \mid x, z) = \sum_{t \in \mathcal{T}} \mu(x, \tilde{z}, t) P(t \mid x, z),$$

where the last equality follows by quasi-random assignment. Furthermore, for each value of $t \in \mathcal{T}$ and $z \in \mathcal{Z}$, $\mu(x, z, t)$ is bounded by

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

Therefore, for a given $z \in \mathcal{Z}$, the sharp lower and upper bounds on $\mu(x, z, t)$ are

$$\begin{aligned}
&\mathbb{E}\left[\mu_1(X, \tilde{z}, T)\pi_1(X, \tilde{z}, T) \mid X = x, Z = z\right] \leq \mu(x, z), \\
&\mu(x, z) \leq \mathbb{E}\left[\mu_1(X, \tilde{z}, T)\pi_1(X, \tilde{z}, T) + K\pi_0(X, \tilde{z}, T) \mid X = x, Z = z\right]
\end{aligned}$$

for any $\tilde{z} \in \mathcal{Z}$. Since $\mu_1(x, z)\pi_1(x, z)$ is observed, this implies bounds on $\mu_0(x, z)\pi_0(x, z)$. Assuming $\pi_0(x, z) > 0$, this implies a bound on $\mu_0(x, z)$ since $\pi_0(x, z)$ is also observed.

E.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. I now show how $\underline{\mathcal{E}}$ can be translated into an equivalent reduction in ex-post errors $P(C = 1, Y^* = 1)$ 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 (8). By the definition of approximate expected utility maximization, $\underline{\mathcal{E}}$ is an upper bound on

$$\mathbb{E}[u(C^*(X, V), Y^*; X_0) - u(C, Y^*; X_0)] = \sum_{x_0 \in \mathcal{X}_0} \left\{|u_{0,1}(x_0)|\Delta_{0,0}(x_0) + |u_{1,1}(x_0)|\Delta_{1,1}(x_0)\right\} P(x_0), \tag{26}$$

where $C^*(X, V)$ is expected utility maximization choice at (X, V), $\Delta_{0,0}(x_0) = \mathbb{E}[(1 - C)(1 - Y_1^*) - (1 - C^*(X, V))(1 - Y_1^*) \mid X_0 = x_0]$ is the reduction of ex-post errors that select $C = 0$ when Y_1^* is small, and $\Delta_{1,1}(x_0) = \mathbb{E}[CY_1^* - C^*(X, V)Y_1^* \mid X_0 = x_0]$ 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 $x_0 \in \mathcal{X}_0$,

$$\max_{\tilde{x}_1 \in \mathcal{X}_1} \{\mu_1(x_0, \tilde{x}_1) - \overline{\epsilon}(x_0, \tilde{x}_1)\} \leq |u_{0,1}(x_0)| \leq \min_{\tilde{x}_1 \in \mathcal{X}_1} \max_{\tilde{x}_1 \in \mathcal{X}} \{\overline{\mu}_0(x_0, \tilde{x}_1) - \overline{\epsilon}(x_0, \tilde{x}_1)\}, \tag{27}$$

and define $|u_{0,1}(x_0)| = \max_{\tilde{x}_1 \in \mathcal{X}_1} \{\mu_1(x_0, \tilde{x}_1) - \overline{\epsilon}(x_0, \tilde{x}_1)\}$, $|u_{1,1}(x_0)| = 1 - |u_{0,1}(x_0)|$. At this linear utility function in the identified set for expected utility costs $\overline{\epsilon}(x)$, we can calculate the implied reduction in ex-post errors $\Delta_{1,1}(x_0)$ that are equivalent to $\underline{\mathcal{E}}$ in an expected utility sense by calculating

$$\begin{aligned} \max_{\Delta_{0,0}(x_0), \Delta_{1,1}(x_0)} & \sum_{x_0} \Delta_{1,1}(x_0) P(x_0) \\ \text{s.t. } & 0 \leq \Delta_{1,1}(x_0) \leq \mathbb{E}[CY^* \mid X_0 = x_0] \text{ for all } x_0 \in \mathcal{X}_0, \\ & 0 \leq \Delta_{0,0}(x_0) \leq \pi_0(x_0) \text{ for all } x_0 \in \mathcal{X}_0, \\ & \sum_{x_0 \in \mathcal{X}_0} \{|u_{0,1}(x_0)| \Delta_{0,0}(x_0) + |u_{1,1}(x_0)| \Delta_{1,1}(x_0)\} P(x_0) \leq \underline{\mathcal{E}}. \end{aligned} \tag{28}$$

The first constraint imposes that the reduction in ex-post errors $\Delta_{1,1}(x_0)$ must be weakly positive, and can be no greater than the observed ex-post errors $\mathbb{E}[CY^* \mid X_0 = x_0]$ at the decision maker's choices. The second constraint imposes that the reduction in ex-post errors $\Delta_{0,0}(x_0)$ must also be weakly positive, and can be no greater than the observed probability the decision maker selected $C = 0$. This is an upper bound on $\mathbb{E}[(1-C)(1-Y^*) \mid X_0 = x_0]$. The final constraint imposes that the implied expected utility of the change in ex-post errors must be consistent with the expected utility cost $\underline{\mathcal{E}}$.

E.3 Mixed-integer linear program for the share of systematic prediction mistakes

I now show that the share of systematic prediction mistakes in the decision maker's choices $\underline{\lambda}$ defined in (11) 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 that

$$\begin{aligned} \underline{\lambda} := \min_{\omega(x), \epsilon(x)} & \sum_{x} P(x) \omega(x) \text{ s.t.} \\ & \mu_1(x) - \overline{\mu}_0(x') \epsilon(x) - \epsilon(x') \leq 0 \text{ for all pairs } x = (x_0, x_1), x' = (x_0, x_1'), \\ & 0 \leq \epsilon(x) \leq M \cdot \omega(x), \, \omega(x) \in \{0, 1\} \end{aligned}$$

since $\max_{x, x' \in \mathcal{X}} \{\mu_1(x) - \overline{\mu}_0(x')\} \leq 2K$ because $Y_k^* \in [0, 1]$ for all $k = 1, \ldots, K$. Mixed-integer linear programs also appear in several, unrelated econometric problems such as the computation of the maximum score estimator (Florios and Skouras, 2008), and the calculation of empirical welfare maximizing policy rules (Kitagawa and Tetenov, 2018; Viviano, 2020).

E.4 Expected social welfare: identification and inference

E.4.1 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 $\pi_1^*(x)$ as given in (17), which can be rewritten as

$$\theta(x) = \ell(x; \pi_1^*, u^*) P(1 \mid x) + \beta(x; \pi_1^*, u^*) \tag{29}$$

for $\ell(x; \pi_1^*, u^*) := u_{1,1}^* \pi_1^*(x) - u_{0,1}^*(1 - \pi_1^*(x))$ and $\beta(x; \pi_1^*, u^*) := u_{0,1}^*(1 - \pi_1^*(x))$. Total expected social welfare then equals

$$\theta(\pi_1^*, u^*) = \beta(\pi_1^*, u^*) + \sum_{x \in \mathcal{X}} P(x) \ell(x; \pi_1^*, u^*) P(1 \mid x), \tag{30}$$

where $\beta(\pi_1^*, u^*) := \sum_{x \in \mathcal{X}} P(x) \beta(x; \pi_1^*, u^*)$. 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 E.1. *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 $\pi_1^*(x)$. The sharp identified set of total expected social welfare, denoted $\mathcal{H}_P(\theta(\pi_1^*, u^*); \mathcal{B})$, is an interval with $\mathcal{H}_P(\theta(\pi_1^*, u^*); \mathcal{B}) = \left[\underline{\theta}(\pi_1^*, u^*), \overline{\theta}(\pi_1^*, u^*)\right]$, where*

$$\underline{\theta}(\pi_1^*, u^*) = \beta(\pi_1^*, u^*) + \left\{ \min_{\left\{\widetilde{P}(\cdot|x):\, x \in \mathcal{X}\right\}} \sum_{x \in \mathcal{X}} P(x) \ell(x; \pi_1^*, u^*) \widetilde{P}(1 \mid x) \text{ s.t. } \widetilde{P}(\cdot \mid x) \in \mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_{0,x}) \; \forall x \in \mathcal{X} \right\},$$

$$\overline{\theta}(\pi_1^*, u^*) = \beta(\pi_1^*, u^*) + \left\{ \max_{\left\{\widetilde{P}(\cdot|x):\, x \in \mathcal{X}\right\}} \sum_{x \in \mathcal{X}} P(x) \ell(x; \pi_1^*, u^*) \widetilde{P}(1 \mid x) \text{ s.t. } \widetilde{P}(\cdot \mid x) \in \mathcal{H}_P(P(\cdot \mid x); \mathcal{B}_{0,x}) \; \forall x \in \mathcal{X} \right\}.$$

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 E.1 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 E.2. *Under the same set-up as Proposition E.1, conditional on the characteristics X, testing the null hypothesis $H_0 \colon \theta(\pi_1^*, u^*) = \theta_0$ is equivalent to testing whether*

$$\exists \delta \in \mathbb{R}^{d_x - 1} \text{ s.t. } \widetilde{A}_{(\cdot,1)} \left(\theta_0 - \ell^\intercal(\pi_1^*, u^*) P^{c=1, y_1^*=1} - \beta(\pi_1^*, u^*)\right) + \widetilde{A}_{(\cdot,-1)} \delta \leq \begin{pmatrix} -\underline{P}^{c=0, y_1^*=1} \\ \overline{P}^{c=0, y_1^*=1} \end{pmatrix},$$

where $d_x := |\mathcal{X}|$, $\ell(\pi_1^*, u^*)$ *is the* d_x*-dimensional vector with elements* $P(x)\ell(x; \pi_1^*, u^*)$, $P^{c=1,y_1^*=1}$ *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.*[37]

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 (2019) and Cox and Shi (2020) develop inference procedures that exploit the additional linear structure and are valid in low-dimensional settings.

E.4.2 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

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

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 E.3. *Under the same set-up as Proposition E.1, the sharp identified set of total expected social welfare under the decision maker's observed choices, denoted* $\mathcal{H}_P(\theta^{DM}(u^*); \mathcal{B})$*, is an interval with* $\mathcal{H}_P(\theta^{DM}(u^*); \mathcal{B}) = \left[\underline{\theta}^{DM}(u^*), \overline{\theta}^{DM}(u^*)\right]$*, where*

$$\begin{aligned}\underline{\theta}^{DM}(u^*) &= u_{1,1}^* P(C = 1, Y_1^* = 1) + u_{0,1}^* P(C = 0) - u_{0,1}^* \overline{P}(C = 0, Y_1^* = 1) \\ \overline{\theta}^{DM}(u^*) &= u_{1,1}^* P(C = 1, Y_1^* = 1) + u_{0,1}^* P(C = 0) - u_{0,1}^* \underline{P}(C = 0, Y_1^* = 1),\end{aligned}$$

where

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

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

As in Appendix **??**, the bounds $\mathcal{B}_x$ for a binary outcome are an interval, and so Proposition E.1 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

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

system of moment inequalities with a large number of nuisance parameters that enter the moments linearly.

Proposition E.4. *Under the same set-up as Proposition E.1, conditional on the characteristics X, testing the null hypothesis $H_0\colon \theta^{DM}(u^*) = \theta_0$ is equivalent to testing whether*

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

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.

E.5 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.[38]

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 $\pi_1^*(x) \in [0,1]$ denote the probability the policymaker selects $C=1$ given $X=x$. At each $x \in \mathcal{X}$, the policymaker then chooses $\pi_1^*(x)$ to maximize

$$\begin{aligned} &\min_{\widetilde{P}(1|x)} \pi_1^*(x)\widetilde{P}(1 \mid x)u^*_{1,1} + (1-\pi_1^*(x))(1-\widetilde{P}(1 \mid x))u^*_{0,1} \\ &\text{s.t. } \underline{P}(1 \mid x) \leq \widetilde{P}(1 \mid x) \leq \overline{P}(1 \mid x). \end{aligned}$$

Proposition E.5. *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 $\pi_1^*(x) \in [0,1]$ 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*

$$\pi_1^*(x) = \begin{cases} 1 \text{ if } \overline{P}_{Y^*}(1 \mid x) \leq \tau^*, \\ 0 \text{ if } \underline{P}_{Y^*}(1 \mid x) \geq \tau^*, \\ \tau^* \text{ if } \underline{P}_{Y^*}(1 \mid x) < \tau^* < \overline{P}_{Y^*}(1 \mid x). \end{cases}$$

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 payoff relevant characteristics X_0 and each decile of predicted

[38] Rambachan et al. (2021) refer to this as the "first-best problem" in their analysis of algorithmic decision rules.

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.

E.6 Proofs of additional results for the econometric framework

E.6.1 Proof of Proposition E.1

Proof. The researcher's bounds on the unobserved conditional outcome probabilities implies bounds on $\widetilde{P}(1 \mid x) \in \mathcal{H}_P(P(1 \mid x); \mathcal{B}_x)$ as discussed in Section 2.2 of the main text. The result then immediately follows from (30), taking the maximum and minimum over $P(1 \mid x)$ that are consistent with the researcher's bounds. □

E.6.2 Proof of Proposition E.2

Proof. First, rewrite $\theta(\pi_1^*, u^*)$ as

$$\beta(\pi_1^*, u^*) + \ell^\intercal(\pi_1^*, u^*) P_{C,Y_1^*}(1,1) + \ell^\intercal(\pi_1^*, u^*) P_{C,Y_1^*}(0,1),$$

where $\ell^\intercal(\pi_1^*, u^*)$ 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 $H_0 : \theta(\pi_1^*, u^*) = \theta_0$ is equivalent to the null hypothesis that there exists $\widetilde{P}_{C,Y_1^*}(0,1)$ satisfying

$$\ell^\intercal(\pi_1^*, u^*) \widetilde{P}_{C,Y_1^*}(0,1) = \theta(\pi_1^*, u^*) - \beta(\pi_1^*, u^*) - \ell^\intercal(\pi_1^*, u^*) P_{C,Y_1^*}(1,1)$$
$$\underline{P}(C = 0, Y_1^* = 1 \mid X = x) \leq \widetilde{P}(C = 0, Y_1^* = 1 \mid X = x) \leq \overline{P}(C = 0, Y_1^* = 1 \mid X = x) \text{ for all } x \in \mathcal{X}.$$

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 $H_0 : \theta(\pi_1^*, u^*) = \theta_0$ is equivalent to the null hypothesis

$$\exists \widetilde{P}_{C,Y_1^*}(0,1) \text{ satisfying } \ell^\intercal(\pi_1^*, u^*) \widetilde{P}_{C,Y_1^*}(0,1) = \theta_0 - \beta(\pi_1^*, u^*) - \ell^\intercal(\pi_1^*, u^*) P_{C,Y_1^*}(1,1) \text{ and}$$
$$A\widetilde{P}_{C,Y_1^*}(0,1) \leq \begin{pmatrix} -\underline{P}_{C,Y_1^*}(0,1) \\ \overline{P}_{C,Y_1^*}(0,1) \end{pmatrix}.$$

Next, we apply a change of basis argument. Define the full rank matrix Γ, whose first row is equal to $\ell^\intercal(\pi_1^*, u^*)$. The null hypothesis $H_0 : \theta(\pi_1^*, u^*) = \theta_0$ 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 $\Gamma\widetilde{P}_{C,Y_1^*}(0,1) = \begin{pmatrix} \Gamma_{(1,\cdot)}\widetilde{P}_{C,Y_1^*}(0,1) \\ \Gamma_{(-1,\cdot)}\widetilde{P}_{C,Y_1^*}(0,1) \end{pmatrix} = \begin{pmatrix} \theta_0 - \beta(\pi_1^*, u^*) - \ell^\intercal(\pi_1^*, u^*) P_{C,Y_1^*}(1,1) \\ \delta \end{pmatrix}$ 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 alge-

bra. □

E.6.3 Proof of Proposition E.3

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

E.6.4 Proof of Proposition E.4

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 $\widetilde{P}(C=0, Y_1^* \mid X=xx)$. 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. □

E.6.5 Proof of Proposition E.5

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 $\pi_1^*(x) = 1$.

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 $\pi_1^*(x) = 0$.

Case 3: Suppose $\underline{P}(Y_1^* = 1 \mid X = x) < \tau^* < \overline{P}(Y_1^* = 1 \mid X = x)$. First, notice $\pi_1^*(x) = \tau^*$ 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 $\pi_1^*(x)$, expected social welfare equals

$$\pi_1^*(x)P(Y_1^* = 1 \mid X = x)u_{1,1}^* + (1 - \pi_1^*(x))P(Y_1^* = 0 \mid X = x)u_{0,1}^*.$$

The derivative with respect to $P(Y_1^* = 1 \mid X = x)$ equals $\pi_1^*(x)u_{1,1}^* - (1 - \pi_1^*(x))u_{0,1}^*$, which equals zero if $\pi_1^*(x) = \tau^*$. Moreover, worst case expected social welfare at $\pi_1^*(x) = \tau^*$ 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 $\pi_1^*(x)$ delivers strictly lower worst-case expected social welfare in this case.

Consider any $\pi_1^*(x) < \tau^*$. 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 $\pi_1^*(x)$ 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

$$\pi_1^*(x)\underline{P}(Y_1^* = 1 \mid X = x)u_{1,1}^* + (1 - \pi_1^*(x))(1 - \underline{P}(Y_1^* = 1 \mid X = x))u_{0,1}^* <$$

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

Therefore, worst-case expected social welfare for any $\pi_1^*(x) < \tau^*$ is strictly less than worst-case expected social welfare at $\pi_1^*(x) = \tau^*$.

Consider any $\pi_1^*(x) > \tau^*$. 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 $\pi_1^*(w, x)$ 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

$$\pi_1^*(x)\overline{P}(Y_1^* = 1 \mid X = x)u_{1,1}^* + (1 - \pi_1^*(x))(1 - \overline{P}(Y_1^* = 1 \mid X = x))u_{0,1}^* <$$

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

Therefore, worst-case expected social welfare for any $\pi_1^*(x) > \tau^*$ is strictly less than worst-case expected social welfare at $\pi_1^*(x) = \tau^*$. □