Identifying Prediction Mistakes in Observational Data*

Ashesh Rambachan†

Job Market Paper

September 21, 2021

[Link to Most Recent Version]
[Link to Supplementary Materials]

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, pretrial release, and medical testing. I derive a statistical test for whether the decision maker makes systematic prediction mistakes under these assumptions and show how supervised machine learning based methods can be used to apply this test. I provide methods for conducting inference on the ways in which the decision maker's predictions are systematically biased. As an illustration, I apply this econometric framework to analyze the pretrial release decisions of judges in New York City, and I estimate that at least 20% of judges make systematic prediction mistakes about failure to appear risk given defendant characteristics.

*First version: June 2021. I am especially grateful to Isaiah Andrews, Sendhil Mullainathan, Neil Shephard, Elie Tamer and Jens Ludwig for their invaluable feedback, support, and advice. I thank Alex Albright, Nano Barahona, Laura Blattner, Iavor Bojinov, Raj Chetty, Bo Cowgill, Will Dobbie, Xavier Gabaix, Matthew Gentzkow, Ed Glaeser, Yannai Gonczarowski, Larry Katz, Ross Mattheis, Robert Minton, Ljubica Ristovska, Jonathan Roth, Joshua Schwartzstein, Jesse Shapiro, and participants at the Brookings Institution's Artificial Intelligence Conference for many useful comments and suggestions. I 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.

†Harvard University, Department of Economics: asheshr@g.harvard.edu

1 Introduction

Decision makers, such as doctors, judges, and managers, are commonly tasked with making consequential choices 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 detailed information such as the defendant's current criminal charge and prior criminal record. Are these decision makers making systematic prediction mistakes based on the available information? If so, in what ways are their predictions systematically biased? These foundational questions in behavioral economics and psychology (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 in economics and computer science evaluate decision makers' implicit predictions through comparisons of their choices against those made by predictive models.[2] Uncovering systematic prediction mistakes from decisions is challenging, however, 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. Hence, the decision maker's choices may 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 this empirical literature recognizes these challenges (e.g., Kleinberg et al., 2018*a*; Mullainathan and Obermeyer, 2020), it lacks a unifying econometric framework for analyzing the decision maker's choices under as weak as possible 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 predictions are systematically bi-

[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 Einav, Jenkins and Levin (2013), Fuster et al. (2018), Gillis (2019), Dobbie et al. (2020), and Blattner and Nelson (2021) in economics, and see Khandani, Kim and Lo (2010), Hardt, Price and Srebro (2016), Liu et al. (2018), and Coston, Rambachan and Chouldechova (2021) in computer science. For discussions of workforce analytics and resume screening software, see Autor and Scarborough (2008), Jacob and Lefgren (2008), Rockoff et al. (2011), Feldman et al. (2015), Hoffman, Kahn and Li (2018), Erel et al. (2019), Li, Raymond and Bergman (2020), Raghavan et al. (2020), and Frankel (2021).

[2] See, for example, Kleinberg et al. (2015), Berk, Sorenson and Barnes (2016), Chalfin et al. (2016), Chouldechova et al. (2018), Cowgill (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*a*), and Mullainathan and Obermeyer (2020). 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.

ased. 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 inferences about systematic prediction mistakes. I consider empirical settings, such as pretrial release, medical treatment or testing, 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 (Heckman, 1974; Rubin, 1976; Heckman, 1979; Manski, 1989): we only observe 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).[3]

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 derive a statistical test for whether the decision maker makes systematic prediction mistakes and show how machine learning based models can be used to apply this test. Third, I provide methods for conducting inference on the ways in which the decision maker's predictions are systematically biased. These contributions provide, to my knowledge, the first microfounded econometric framework for studying systematic prediction mistakes in these empirical settings, enabling researchers to answer a wider array of behavioral questions under weaker assumptions than existing empirical research.[4] Finally, I apply this econometric framework to analyze the pretrial release decisions of judges in New York City.

I explore the restrictions imposed on the decision maker's choices by expected utility maximization behavior, where the decision maker maximizes some (unknown to the researcher) utility function at beliefs about the outcome given the characteristics as well as some private information.[5,6,7] Due to the missing data problem, the true conditional distribution of the outcome given

[3] A large literature explores whether forecasters, households, or individuals have rational expectations in settings where both outcomes and subjective expectations are observed. See, for example, Manski (2004), Elliott, Timmerman and Komunjer (2005), Elliott, Komunjer and Timmerman (2008), Gennaioli, Ma and Shleifer (2016), Bordalo et al. (2020), D'Haultfoeuille, Gaillac and Maurel (2020). I focus on settings in which we only observe an individual's discrete choices and partial information about the outcome.

[4] Appendix B provides a step-by-step user's guide for empirical researchers interested in applying these methods.

[5] Mourifie, Henry and Meango (2019) and Henry, Meango and Mourifie (2020) analyze the testable implications of Roy-style and extended Roy-style selection. See Heckman and Vytlacil (2006) for an econometric review of Roy selection models. The expected utility maximization model can be interpreted as a generalized Roy model, and I discuss these connections in Section 2.

[6] A literature in decision theory explores conditions under which a decision maker's random choice rule, which summarizes her choice probabilities in each possible menu of actions, has a random utility model representation. See, for example, Gul and Pesendorfer (2006), Gul, Natenzon and Pesendorfer (2014), Lu (2016), and Natenzon (2019). I consider empirical settings in which we only observe choice probabilities from a single menu.

[7] Kubler, Selden and Wei (2014), Echenique and Saito (2015), Chambers, Liu and Martinez (2016), Polisson, Quah and Renou (2020), and Echenique, Saito and Imai (2021) use revealed preference analysis to characterize expected utility maximization behavior in consumer demand settings, in which a consumer's state-contingent consumption

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 preferences in a researcher-specified class that rationalizes observed choices under this model, I therefore say the decision maker is making systematic prediction mistakes based on the characteristics of individuals.

I derive a sharp characterization, based on revealed preference inequalities, of the identified set of utility functions at which the decision maker's choices are consistent with expected utility maximization behavior at accurate beliefs. If these revealed preference inequalities are satisfied at a candidate utility function, then some distribution of private information can be constructed such that the decision maker cannot do better than her observed choices in an expected utility sense.[8] If the identified set of utility functions is empty, then the decision maker is making systematic prediction mistakes as there is no combination of preferences and private information at which her observed choices are consistent with expected utility maximization at accurate beliefs.

I prove that without further assumptions systematic prediction mistakes are *untestable*. If either all characteristics of individuals directly affect the decision maker's utility function or the missing data can take any value, then the identified set of utility functions is non-empty. Any variation in the decision maker's conditional choice probabilities can be rationalized by a utility function and private information that sufficiently vary across all the characteristics. However, placing an exclusion restriction on which characteristics may directly affect utility and constructing informative bounds on the missing data together restore the testability of expected utility maximization behavior.[9] Under such an exclusion restriction, variation in the decision maker's choices across characteristics that do not directly affect utility 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 preferences and econometric assumptions to address the missing data problem are therefore sufficient to iden-

choices across several budget sets are observed. See the review in Echenique (2020).

[8]By searching for any distribution of private information that rationalizes the decision maker's choices, my analysis follows in the spirit of the robust information design literature (e.g., Kamenica and Gentzkow, 2011; Bergemannn and Morris, 2013, 2016, 2019; Kamenica, 2019). Syrgkanis, Tamer and Ziani (2018) and Bergemann, Brooks and Morris (2019) use results from this literature to study multiplayer games, whereas I analyze the choices of a single decision maker. Gualdani and Sinha (2020) also analyzes single-agent settings under weak assumptions on the information environment.

[9]The exclusion restriction on which characteristics may directly affect utility complements recent results on the validity of "marginal outcome tests" for discrimination (Bohren et al., 2020; Canay, Mogstad and Mountjoy, 2020; Gelbach, 2021; Hull, 2021). In the special case of a binary decision and binary outcome, the expected utility maximization model is a generalization of the extended Roy model analyzed in Canay, Mogstad and Mountjoy (2020) and Hull (2021). I formalize these connections in Section 3.

tify systematic prediction mistakes.

These results clarify what conclusions can be logically drawn about systematic prediction mistakes given a set of researcher-specified assumptions on the decision maker's preferences and the missing data. These testable restrictions arise from the *joint* null hypothesis that the decision maker maximizes expected utility at accurate beliefs about the outcome given the characteristics and that their preferences satisfy the conjectured exclusion restriction.[10] Therefore, if these restrictions are satisfied, we cannot logically reject that the decision maker's choices maximize expected utility at some preferences in the researcher's class and beliefs about the outcomes given the characteristics that lie in the identified set. Stronger conclusions would require stronger assumptions on preferences or tighter bounds on the missing data.

I show that testing these restrictions implied by expected utility maximization at accurate beliefs is equivalent to a moment inequality problem. This is a well-studied problem in econometrics for which many procedures are available (e.g., see the reviews by Canay and Shaikh, 2017; Molinari, 2020). The number of moment inequalities grows with the dimensionality of the observable characteristics of individuals, which will typically be quite large in empirical applications. To deal with this practical challenge, I discuss how supervised machine learning methods may be used to reduce the dimension of this testing problem. Researchers may construct a prediction function for the outcome on held out data and partition the characteristics into percentiles of predicted risk based on this estimated prediction function. Testing implied revealed preference inequalities across percentiles of predicted risk is a valid test of the joint null hypothesis that the decision maker's choices maximize expected utility at preferences satisfying the conjectured exclusion restriction and accurate beliefs. This provides, to my knowledge, the first microfounded procedure for using supervised machine learning based prediction functions to identify systematic prediction mistakes.

With this framework in place, I further establish that the data are informative about how the decision maker's predictions are systematically biased. I extend the behavioral model to 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.[11] 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; Gabaix, 2014; Caplin and Dean, 2015; Bordalo et al., 2016; Handel and Schwartzstein,

[10]This finding echoes a classic insight in finance that testing whether variation in asset prices reflect violations of rational expectations requires assumptions about admissible variation in underlying stochastic discount factors. See Campbell (2003), Cochrane (2011) for reviews and Augenblick and Lazarus (2020) for a recent contribution.

[11]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.

2018). 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. These bounds again arise because any variation in the decision maker's choices across characteristics that do not directly affect utility must only arise due to variation in beliefs. Comparing observed variation in the decision maker's choice probabilities against possible variation in the probability of the outcome is therefore informative about the extent to which the decision maker's beliefs are inaccurate.

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.[12] For each judge, I observe the conditional probability that she releases a defendant given a rich set of characteristics (e.g., race, age, current charge, prior criminal record, etc.) as well as the conditional probability that a released defendant fails to appear in court. The conditional failure to appear rate among detained defendants is unobserved due to the missing data problem. If all defendant characteristics may directly affect the judge's preferences or the conditional failure to appear rate among detained defendants may take any value, then my theoretical results establish that the judge's choices are always consistent with expected utility maximization behavior at accurate beliefs. Without further assumptions, we cannot logically rule out that the judge's choices reflect either preferences that vary richly based on defendant characteristics or sufficiently predictive private information.

However, empirical researchers often assume that while judges may engage in taste-based discrimination on a defendant's race, a defendant's prior criminal record only affects their 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 on the judge's preferences and quasi-experimental bounds on the missing data, 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 observed failure to appear rate than the researcher's upper bound on the failure to appear rate of all detained defendants? If not, there is no combination of preferences satisfying the conjectured exclusion restriction nor private information such that the judge's choices maximize expected utility at accurate beliefs about failure

[12] Several empirical papers also study the New York City pretrial release system. Leslie and Pope (2017) estimates the effects of pretrial detention on criminal case outcomes. Arnold, Dobbie and Hull (2020*b*) and Arnold, Dobbie and Hull (2020*a*) estimate whether judges and pretrial risk assessments respectively discriminate against black defendants. Kleinberg et al. (2018*a*) studies whether a statistical risk assessment could improve pretrial outcomes in New York City. I discuss the differences between my analysis and this prior research in Section 5.

to appear risk given defendant characteristics.

By testing for such misrankings in their pretrial release decisions, 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 on preferences 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.

Finally, to highlight broader 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 show that expected social welfare under a candidate decision rule is partially identified, and inference on counterfactual expected social welfare can be reduced to testing moment inequalities with nuisance parameters that enter the moments linearly (e.g., see Gafarov, 2019; Andrews, Roth and Pakes, 2019; Cho and Russell, 2020; Cox and Shi, 2020). Using these results, 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.[13] 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. More broadly, designing algorithmic decision

[13]This finding relates to a computer science literature on "human-in-the-loop" analyses of algorithmic decision support systems (e.g., Tan et al., 2018; Green and Chen, 2019*a*,*b*; De-Arteaga, Fogliato and Chouldechova, 2020; Hilgard et al., 2021). Recent methods estimate whether a decision should be automated by an algorithm or instead be deferred to an existing decision maker (Madras, Pitassi and Zemel, 2018; Raghu et al., 2019; Wilder, Horvitz and Kamar, 2020; De-Arteaga, Dubrawski and Chouldechova, 2021). I show that understanding whether to automate or defer requires assessing whether the decision maker makes systematic prediction mistakes.

rules requires carefully assessing their predictive accuracy and their effects on disparities across groups (e.g., Barocas, Hardt and Narayanan, 2019; Mitchell et al., 2019; Chouldechova and Roth, 2020). These findings suggest that it is also essential to analyze whether the existing decision makers make systematic prediction mistakes and if so, on what decisions.

This paper relates to a large empirical literature that evaluates decision makers' implicit predictions through either comparisons of their choices against those made by machine learning based models or estimating structural models of decision making in particular empirical settings. While the challenges of unknown preferences and information sets are recognized, researchers typically resort to strong assumptions. Kleinberg et al. (2018*a*) and Mullainathan and Obermeyer (2020) restrict preferences to be constant across both decisions and decision makers. Lakkaraju and Rudin (2017), Chouldechova et al. (2018), Coston et al. (2020), and Jung et al. (2020*a*) assume that observed choices were as-good-as randomly assigned given the characteristics, eliminating the problem of unknown information sets. Recent work introduces parametric models for the decision maker's private information, such as Abaluck et al. (2016), Arnold, Dobbie and Hull (2020*b*), Chan, Gentzkow and Yu (2020), and Jung et al. (2020*b*). See also Currie and Macleod (2017), Ribers and Ullrich (2020), and Marquardt (2021). I develop an econometric framework for studying systematic prediction mistakes that only requires exclusion restrictions on which characteristics affect the decision maker's preferences but no further restrictions. I model the decision maker's information environment fully nonparametrically. This enables researchers to study systematic prediction mistakes in many empirical settings under weaker assumptions than existing research.

My identification results build on a literature that derives the testable implications of various behavioral models in "state-dependent stochastic choice (SDSC) data" (Caplin and Martin, 2015; Caplin and Dean, 2015; Caplin, 2016; Caplin et al., 2020; Caplin and Martin, 2021). While useful in analyzing lab-based experiments, such characterization results have limited applicability due to the difficulty of collecting such SDSC data (Gabaix, 2019; Rehbeck, 2020). I focus on common empirical settings in which the data suffer from a missing data problem, and I 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 experiments, 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 outcomes, multi-valued outcomes, and multiple choices. Lu (2019) shows that a decision maker's state-dependent utilities and beliefs can be identified provided choice probabilities across multiple informational treatments are observed.

2 An Empirical Model of Expected Utility Maximization

A decision maker makes choices for many individuals based on a prediction of an unknown outcome using each individual's characteristics. Under what conditions do the decision maker's choices maximize expected utility at some (unknown to us) utility function and accurate beliefs given the characteristics and some private information?

2.1 Setting and Observable Data

A decision maker makes choices for many individuals based on predictions of an unknown outcome. The decision maker selects a binary choice $c \in \{0,1\}$ for each individual. Each individual is summarized by characteristics $(w,x) \in \mathcal{W} \times \mathcal{X}$ and potential outcomes $\vec{y} := (y_0, y_1)$.[14] The *potential outcome* $y_c \in \mathcal{Y}$ is the outcome that would occur if the decision maker were to select choice $c \in \{0,1\}$. The characteristics and potential outcomes have finite support, and I denote $d_w := |\mathcal{W}|, d_x := |\mathcal{X}|$. The random vector $(W, X, C, \vec{Y}) \sim P$ defined over $\mathcal{W} \times \mathcal{X} \times \{0,1\} \times \mathcal{Y}^2$ summarizes the joint distribution of the characteristics, the decision maker's choices, and the potential outcomes across all individuals. I assume $P(W = w, X = x) \geq \delta$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$ for some $\delta > 0$ throughout the paper.

The researcher observes the characteristics of each individual as well as the decision maker's choice. There is, however, a *missing data problem*: the researcher only observes the potential outcome associated with the choice selected by the decision maker (Rubin, 1974; Holland, 1986). Defining the observable outcome as $Y := CY_1 + (1-C)Y_0$, the researcher therefore observes the joint distribution $(W, X, C, Y) \sim P$. I assume the researcher knows this population distribution with certainty in order to focus on the identification challenges in this setting. The researcher observes the decision maker's *conditional choice probabilities*

$$\pi_c(w,x) := P(C = c \mid W = w, X = x) \text{ for all } (w,x) \in \mathcal{W} \times \mathcal{X}$$

as well as the conditional potential outcome probabilities

$$P_{Y_c}(y_c \mid c, w, x) := P(Y_c = y_c \mid C = c, W = w, X = x) \text{ for all } (w,x) \in \mathcal{W} \times \mathcal{X}, c \in \{0,1\}.$$

The counterfactual potential outcome probabilities $P(Y_0 = y_0 \mid C = 1, W = w, X = x)$, $P(Y_1 = y_1 \mid C = 0, W = w, X = x)$ are not observed due to the missing data problem.[15]

[14] The characteristics $(w,x) \in \mathcal{W} \times \mathcal{X}$ will play different roles in the expected utility maximization model, and so I introduce separate notation now.

[15] I adopt the convention that $P(\vec{Y} = \vec{y} \mid C = c, W = w, X = x) = 0$ if $P(C = c \mid W = w, X = x) = 0$.

As a consequence, the *choice-dependent potential outcome probabilities*

$$P_{\vec{Y}}(\vec{y} \mid c, w, x) := P(\vec{Y} = \vec{y} \mid C = c, W = w, X = x)$$

are unobserved for all $(w, x) \in \mathcal{W} \times \mathcal{X}$, $c \in \{0, 1\}$.

Notation: 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 $P_{\vec{Y}}(\cdot \mid c, w, x) \in \Delta(\mathcal{Y}^2)$ denote the vector of conditional potential outcome probabilities given $C = c$ and characteristics $(w, x) \in \mathcal{W} \times \mathcal{X}$. For any pair $c, \tilde{c} \in \{0, 1\}$ I write $P_{Y_c}(y_c \mid \tilde{c}, w, x) := P(Y_c = y_c \mid C = \tilde{c}, W = w, X = x)$ and $P_{Y_c}(\cdot \mid \tilde{c}, w, x) \in \Delta(\mathcal{Y})$ as the distribution of the potential outcome Y_c given $C = \tilde{c}$ and characteristics (w, x). Analogously, $P_{\vec{Y}}(\vec{y} \mid w, x) := P(\vec{Y} = \vec{y} \mid W = w, X = x)$ with $P_{\vec{Y}}(\cdot \mid w, x) \in \Delta(\mathcal{Y}^2)$ and $P_{Y_c}(y_c \mid W = w, X = x) := P(Y_c = y_c \mid W = w, X = x)$ with $P_{Y_c}(\cdot \mid w, x) \in \Delta(\mathcal{Y})$ denote the distributions of potential outcomes given the characteristics $(w, x) \in \mathcal{W} \times \mathcal{X}$.

Throughout the paper, I model the researcher's assumptions about the missing data problem in the form of bounds on the choice-dependent potential outcome probabilities.

Assumption 2.1 (Bounds on Missing Data)**.** For each $c \in \{0, 1\}$ and $(w, x) \in \mathcal{W} \times \mathcal{X}$, there exists a known subset $\mathcal{B}_{c,w,x} \subseteq \Delta(\mathcal{Y}^2)$ satisfying

(i) $P_{\vec{Y}}(\cdot \mid c, w, x) \in \mathcal{B}_{c,w,x}$,

(ii) $\sum_{y_{\tilde{c}} \in \mathcal{Y}} \widetilde{P}_{\vec{Y}}(\cdot, y_{\tilde{c}} \mid c, w, x) = P_{Y_c}(\cdot \mid c, w, x)$ for all $\widetilde{P}_{\vec{Y}}(\cdot \mid c, w, x) \in \mathcal{B}_{c,w,x}$.

Let $\mathcal{B}_{w,x}$ denote the collection $\{\mathcal{B}_{0,w,x}, \mathcal{B}_{1,w,x}\}$ and $\mathcal{B}$ denote the collection $\mathcal{B}_{w,x}$ at all $(w, x) \in \mathcal{W} \times \mathcal{X}$.

In some cases, researchers may wish to analyze the decision maker's choices without placing any further assumptions on the missing data, which corresponds to setting $\mathcal{B}_{c,w,x}$ equal to the set of all choice-dependent potential outcome probabilities that are consistent with the data. In other cases, researchers may use quasi-experimental variation or introduce additional assumptions to provide informative bounds on the choice-dependent potential outcome probabilities, as I discuss in Section 3.

Under Assumption 2.1, various features of the joint distribution $(W, X, C, \vec{Y}) \sim P$ are partially identified. The sharp identified set for the distribution of the potential outcome vector given the characteristics $(w, x) \in \mathcal{W} \times \mathcal{X}$, denoted by $\mathcal{H}_P(P_{\vec{Y}}(\cdot \mid w, x); \mathcal{B}_{w,x})$, equals the set of $\widetilde{P}(\cdot \mid w, x) \in \Delta(\mathcal{Y}^2)$ satisfying, for all $\vec{y} \in \mathcal{Y}^2$,

$$\widetilde{P}(\vec{y} \mid w, x) = \widetilde{P}_{\vec{Y}}(\vec{y} \mid 0, w, x)\pi_0(w, x) + \widetilde{P}_{\vec{Y}}(\vec{y} \mid 1, w, x)\pi_1(w, x)$$

for some $\widetilde{P}_{\vec{Y}}(\cdot \mid 0, w, x) \in \mathcal{B}_{0,w,x}$, $\widetilde{P}_{\vec{Y}}(\cdot \mid 1, w, x) \in \mathcal{B}_{1,w,x}$. If the bounds $\mathcal{B}_{0,w,x}, \mathcal{B}_{1,w,x}$ are singletons for all $(w, x) \in \mathcal{W} \times \mathcal{X}$, then the joint distribution $(W, X, C, \vec{Y}) \sim P$ is point identified by the data and the researcher's assumptions.

The main text of the paper assumes that the decision maker faces only two choices and that the characteristics have finite support. My results extend to settings with multiple choices directly, and I provide an extension to the case with continuous characteristics in Supplement F.

2.2 Motivating Empirical Applications

In this section, I illustrate how several motivating empirical applications map into this setting.

Example (Medical Treatment)**.** A doctor decides what medical treatment to give a patient based on a prediction of how that medical treatment will affect the patient's health outcomes (Chandra and Staiger, 2007; Manski, 2017; Currie and Macleod, 2017; Abaluck et al., 2020; Currie and Macleod, 2020). For example, a doctor decides whether to give reperfusion therapy $C \in \{0, 1\}$ to an admitted patient that has suffered from a heart attack (Chandra and Staiger, 2020). The potential outcomes $Y_0, Y_1 \in \{0, 1\}$ denote whether the patient would have died within 30 days of admission had the doctor not given or given reperfusion therapy respectively. The characteristics (W, X) summarize rich information about the patient that is available at the time of admission such as demographic information, collected vital signs and the patient's prior medical history. The researcher observes the doctor's conditional treatment probability $\pi_1(w, x)$, the 30-day mortality rate among patients that received reperfusion therapy $P_{Y_1}(1 \mid 1, w, x)$ and did not receive reperfusion therapy $P_{Y_0}(1 \mid 0, w, x)$. The counterfactual 30-day mortality rates $P_{Y_0}(1 \mid 1, w, x)$ and $P_{Y_1}(1 \mid 0, w, x)$ are not observed. ▲

In a large class of empirical applications, the decision maker's choice does not have a direct causal effect on the outcome of interest, but still generates a missing data problem. In these settings, the potential outcome given $C = 0$ satisfies $Y_0 \equiv 0$, and $Y_1 := Y^*$ is a latent outcome that is revealed whenever the decision maker selects choice $C = 1$. Hence, the observable outcome is $Y := C \cdot Y^*$. These *screening decisions* are a leading class of prediction policy problems (Kleinberg et al., 2015).

Example (Pretrial Release)**.** A judge decides whether to to detain or release defendants $C \in \{0, 1\}$ awaiting trial (Arnold, Dobbie and Yang, 2018; Kleinberg et al., 2018*a*; Arnold, Dobbie and Hull, 2020*b*). The latent outcome $Y^* \in \{0, 1\}$ is whether a defendant would commit pretrial misconduct if released. The characteristics (W, X) summarize information about the defendant that is available at the pretrial release hearing such as demographic information, the current charges filed against the defendant and the defendant's prior criminal record. The researcher observes the characteristics of each defendant, whether the judge released them, and whether the defendant committed pretrial

misconduct only if the judge released them. The judge's conditional release rate $\pi_1(w, x)$ and conditional pretrial misconduct rate among released defendants $P_{Y^*}(1 \mid 1, w, x)$ are observed. The conditional pretrial misconduct rate among detained defendants $P_{Y^*}(1 \mid 0, w, x)$ is unobserved. ▲

Example (Medical Testing and Diagnoses)**.** A doctor decides whether to conduct a costly medical test or make a particular diagnosis (Abaluck et al., 2016; Ribers and Ullrich, 2019; Chan, Gentzkow and Yu, 2020). For example, shortly after an emergency room visit, a doctor decides whether to conduct a stress test on patients $C \in \{0, 1\}$ to determine whether they had a heart attack (Mullainathan and Obermeyer, 2020). The latent outcome $Y^* \in \{0, 1\}$ is whether the patient had a heart attack. The characteristics (W, X) summarize information that is available about the patient such as their demographics, reported symptoms, and prior medical history. The researcher observes the characteristics of each patient, whether the doctor conducted a stress test and whether the patient had a heart attack only if the doctor conducted a stress test. The doctor's conditional stress testing rate $\pi_1(w, x)$ and the conditional heart attack rate among stress tested patients $P_{Y^*}(1 \mid 1, w, x)$ are observed. The conditional heart attack rate among untested patients $P_{Y^*}(1 \mid 0, w, x)$ 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).[16] The latent outcome $Y^* \in \mathcal{Y}$ is some measure of on-the-job productivity, which may be length of tenure since turnover is costly. The characteristics (W, X) are various information about the applicant such as demographics, education level and prior work history. The researcher observes the characteristics of each applicant, whether the manager hired the applicant, and their length of tenure only if hired. The manager's conditional hiring rate $\pi_1(w, x)$ and the conditional distribution of tenure lengths among hired applicants $P_{Y^*}(y^* \mid 1, w, x)$ are observed. The distribution of tenure lengths among rejected applicants $P_{Y^*}(y^* \mid 0, w, x)$ is unobserved. ▲

Other examples of a screening decision include loan approvals (e.g., Fuster et al., 2018; Dobbie et al., 2020; Blattner and Nelson, 2021) and child welfare screenings (Chouldechova et al., 2018).

2.3 Expected Utility Maximization Behavior

I examine the restrictions imposed on the decision maker's choices by expected utility maximization. I define the two main ingredients of the expected utility maximization model. A utility function summarizes the decision maker's preferences over choices, outcomes, and characteristics. Private information is some additional random variable $V \in \mathcal{V}$ that is available to the decision maker but not the researcher, and may be predictive of the outcome.

[16] 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 the outcome $Y^* \in \{0, 1\}$ is whether the applicant is ultimately hired by the firm.

Definition 1. A *utility function* $U\colon \{0,1\}\times\mathcal{Y}^2\times\mathcal{W}\to\mathbb{R}$ specifies the payoff associated with each choice-outcome pair, where $U(c,\vec{y};w)$ is the payoff associated with choice c and potential outcome vector $\vec{y}$ at characteristics $w\in\mathcal{W}$. Let $\mathcal{U}$ denote the feasible set of utility functions specified by the researcher.

Definition 2. The decision maker's *private information* is a random variable $V\in\mathcal{V}$.

Under the model, the decision maker observes the characteristics (W,X) as well as some private information $V\in\mathcal{V}$ prior to selecting a choice. The private information summarizes all additional information that is available to the decision maker but unobserved by the researcher. In empirical settings, researchers often worry that the decision maker observes additional private information that is not recorded in the observable data. For example, in medical treatment, doctors may learn useful information about the patient's current health in an exam. In pretrial release, judges may glean useful information about defendants from courtroom interactions. But these interactions are often not recorded. Since it is unobservable to the researcher, I explore restrictions on the decision maker's behavior without placing distributional assumptions on their private information.

Based on this information set, the decision maker forms beliefs about the unknown outcome and selects a choice to maximize expected utility. The expected utility maximization model is summarized by a joint distribution over the characteristics, private information, choices and the outcome, denoted by $(W,X,C,V,\vec{Y})\sim Q$.

Definition 3. The decision maker's choices are *consistent with expected utility maximization* if there exists a utility function $U\in\mathcal{U}$ and joint distribution $(W,X,V,C,\vec{Y})\sim Q$ satisfying

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

$$\mathbb{E}_Q\left[U(c,\vec{Y};W)\mid W=w,X=x,V=v\right]\geq\mathbb{E}_Q\left[U(c',\vec{Y};W)\mid W=w,X=x,V=v\right].$$

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

iii. Data Consistency: For all $(w,x)\in\mathcal{W}\times\mathcal{X}$, there exists $\widetilde{P}_{\vec{Y}}(\cdot\mid 0,w,x)\in\mathcal{B}_{0,w,x}$ and $\widetilde{P}_{\vec{Y}}(\cdot\mid 1,w,x)\in\mathcal{B}_{1,w,x}$ satisfying for all $\vec{y}\in\mathcal{Y}^2$ and $c\in\{0,1\}$

$$Q(w,x,c,\vec{y})=\widetilde{P}_{\vec{Y}}(\vec{y}\mid c,w,x)P(c,w,x).$$

Definition 4. The *identified set of utility functions*, denoted by $\mathcal{H}_P(U;\mathcal{B})\subseteq\mathcal{U}$, is the set of utility functions $U\in\mathcal{U}$ such that there exists a joint distribution $(W,X,V,C,\vec{Y})\sim Q$ satisfying Definition 3.

In words, the decision maker's choices are consistent with expected utility maximization if three conditions are satisfied. First, if the decision maker selects a choice c with positive probability given $W = w, X = x, V = v$ under the model Q, then it must have been optimal to do so ("Expected Utility Maximization"). The decision maker may flexibly randomize across choices whenever they are indifferent.[17] 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 (W, X, V).[18] Finally, the joint distribution of characteristics, choices and outcomes under the model Q must be consistent with the observable joint distribution P ("Data Consistency").

The key restriction of the expected utility maximization model is that only the characteristics $W \in \mathcal{W}$ directly enter the decision maker's utility function. The decision maker's preferences satisfy an *exclusion restriction* on their private information $V \in \mathcal{V}$ and the characteristics X. The private information $V \in \mathcal{V}$ and characteristics $X \in \mathcal{X}$ may only affect their beliefs. In medical treatment, the utility function specifies the doctor's payoffs from treating a patient given their potential health outcomes. Researchers commonly assume that a doctor's preferences are constant across patients, and patient characteristics only affect beliefs about potential health outcomes under the treatment (e.g., Chandra and Staiger, 2007, 2020).[19] In pretrial release, the utility function specifies a judge's relative payoffs from detaining a defendant that would not commit pretrial misconduct and releasing a defendant that would commit pretrial misconduct. These payoffs may vary based on only some defendant characteristics W. For example, the judge may engage in tasted-based discrimination against black defendants (Becker, 1957; Arnold, Dobbie and Yang, 2018; Arnold, Dobbie and Hull, 2020*b*), 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 preferences. First, as mentioned, exclusion restrictions on the decision maker's preferences are common in existing empirical research. The researcher may therefore appeal to these established modelling choices to guide this assumption. Second, the

[17] Definition 3 requires the decision maker's *fully* maximize expected utility yet researchers may suspect that the decision maker faces some cognitive or computational limitations that prevent them from doing so completely. Appendix C.2 extends the behavioral model to analyze whether the decision maker's choices are consistent with ϵ_w-approximate expected utility maximization, in which the decision maker only approximately maximizes expected utility.

[18] The "Information Set" condition is related to sensitivity analyses in causal inference that assumes there is some unobserved confounder such that the decision is only unconfounded conditional on both the observable characteristics and the unobserved confounder (e.g., Rosenbaum, 2002; Imbens, 2003; Kallus and Zhou, 2018; Yadlowsky et al., 2020). See Supplement G.1 for further discussion of this connection.

[19] Similarly, in medical testing and diagnosis decisions, researchers assume that a doctor's preferences are constant across patients, and patient characteristics only affect beliefs about the probability of an underlying medical condition (e.g., Abaluck et al., 2016; Chan, Gentzkow and Yu, 2020; Mullainathan and Obermeyer, 2020).

exclusion restriction may be normatively motivated, summarizing social or legal restrictions on what observable characteristics ought not to directly enter preferences. Third, the researcher may conduct a sensitivity analysis and report how their analysis of expected utility maximization varies as the choice of payoff-relevant characteristics varies. Such a sensitivity analysis summarizes how flexible the decision maker's preferences must be across observable characteristics in order to rationalize choices.

If Definition 3 is satisfied, then the decision maker's implied beliefs about the outcome given the observable characteristics under the expected utility maximization model, denoted by $Q_{\vec{Y}}(\cdot \mid w, x) \in \Delta(\mathcal{Y}^2)$, lie in the identified set for the distribution of the outcome given the observable characteristics. This is an immediate consequence of Data Consistency in Definition 3.

Lemma 2.1. *If the decision maker's choices are consistent with expected utility maximization, then* $Q_{\vec{Y}}(\cdot \mid w, x) \in \mathcal{H}_P(P_{\vec{Y}}(\cdot \mid w, x); \mathcal{B}_{w,x})$ *for all* $(w, x) \in \mathcal{W} \times \mathcal{X}$.

Therefore, if the decision maker's choices are consistent with expected utility maximization, then their implied beliefs $Q_{\vec{Y}}(\cdot \mid w, x)$ must be "accurate" in this sense. Conversely, if the decision maker's choices are inconsistent with expected utility maximization, then there exists is no configuration of utility function and private information such that their choices would maximize expected utility given any implied beliefs in the identified set for the distribution of the outcome conditional on the characteristics. In this case, their implied beliefs are systematically mistaken.

Definition 5. The decision maker is making *detectable prediction mistakes* based on the observable characteristics if their choices are inconsistent with expected utility maximization, meaning $\mathcal{H}_P(U; \mathcal{B}) = \emptyset$.

I refer to this as a "detectable prediction mistake" as the interpretation of a prediction mistake under Definitions 3-5 is tied to both the researcher-specified bounds on the missing data $\mathcal{B}_{0,w,x}, \mathcal{B}_{1,w,x}$ (Assumption 2.1) and the feasible set of utility functions $\mathcal{U}$ (Definition 1). Less informative bounds on the missing data imply that expected utility maximization places fewer restrictions on behavior as there are more candidate values of the missing choice-dependent potential outcome probabilities that may rationalize choices. Observed behavior that is consistent with expected utility maximization at bounds $\mathcal{B}_{0,w,x}, \mathcal{B}_{1,w,x}$ may, in fact, be inconsistent with expected utility maximization at alternative, tighter bounds $\widetilde{\mathcal{B}}_{0,w,x}, \widetilde{\mathcal{B}}_{1,w,x}$.[20] Analogously, a larger feasible set of utility functions $\mathcal{U}$ implies that expected utility maximization places fewer restrictions on behavior as the researcher

[20]Consider an extreme case in which $P_{\vec{Y}}(\cdot \mid w, x)$ is partially identified under bounds $\mathcal{B}_{0,w,x}, \mathcal{B}_{1,w,x}$ but point identified under alternative bounds $\widetilde{\mathcal{B}}_{0,w,x}, \widetilde{\mathcal{B}}_{1,w,x}$. Under Definitions 3-5, a detectable prediction mistake at bounds $\widetilde{\mathcal{B}}_{0,w,x}, \widetilde{\mathcal{B}}_{1,w,x}$ means that the decision maker's implied beliefs $Q_{\vec{Y}}(\cdot \mid w, x)$ do not equal the point identified quantity $P_{\vec{Y}}(\cdot \mid w, x)$, yet a detectable prediction mistake at bounds $\mathcal{B}_{0,w,x}, \mathcal{B}_{1,w,x}$ means that the decision maker's implied beliefs $Q_{\vec{Y}}(\cdot \mid w, x)$ do not lie in the identified set $\mathcal{H}_P(P_{\vec{Y}}(\cdot \mid w, x); \mathcal{B}_{w,x})$.

is entertaining a larger set of utility functions that may rationalize choices. Definition 5 must therefore be interpreted as a prediction mistake that can be *detected* given the researcher's assumptions on both the missing data and the feasible set of utility functions.

Remark 2.1. 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 (Bohren et al., 2020; Canay, Mogstad and Mountjoy, 2020; Gelbach, 2021; Hull, 2021). Defining the expected benefit functions $\Lambda_0(w,x,v) = \mathbb{E}_Q\left[U(0,\vec{Y};W) \mid W=w, X=x, V=v\right], \Lambda_1(w,x,v) = \mathbb{E}_Q\left[U(1,\vec{Y};W) \mid W=w, X=x, V=v\right]$, the expected utility maximization model is a generalized Roy model that imposes that the observable characteristics $W \in \mathcal{W}$ enter into the utility function and affect beliefs, whereas the observable characteristics $X \in \mathcal{X}$ and private information $V \in \mathcal{V}$ only affect beliefs. The expected utility maximization model also takes no stand on how the decision maker resolves indifferences, and so it is an *incomplete* model of decision making.

2.4 Characterization Result

The decision maker's choices are consistent with expected utility maximization if and only if there exists a utility function $U \in \mathcal{U}$ and values of the missing data that satisfy a series of revealed preference inequalities.

Theorem 2.1. *The decision maker's choices are consistent with expected utility maximization if and only if there exists a utility function $U \in \mathcal{U}$ and $\widetilde{P}_{\vec{Y}}(\cdot \mid 0,w,x) \in \mathcal{B}_{0,w,x}$ and $\widetilde{P}_{\vec{Y}}(\cdot \mid 1,w,x) \in \mathcal{B}_{1,w,x}$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$ satisfying*

$$\mathbb{E}_Q\left[U(c,\vec{Y};W) \mid C=c, W=w, X=x\right] \geq \mathbb{E}_Q\left[U(c',\vec{Y};W) \mid C=c, W=w, X=x\right]. \quad (1)$$

for all $c \in \{0,1\}$, $(w,x) \in \mathcal{W} \times \mathcal{X}$ with $\pi_c(w,x) > 0$ and $c' \neq c$, where the joint distribution $(W,X,C,\vec{Y}) \sim Q$ is given by $Q(w,x,c,\vec{y}) = \widetilde{P}_{\vec{Y}}(\vec{y} \mid c,w,x)P(c,w,x)$.

Corollary 2.1. *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 $\widetilde{P}_{\vec{Y}}(\cdot \mid 0,w,x) \in \mathcal{B}_{0,w,x}$, $\widetilde{P}_{\vec{Y}}(\cdot \mid 1,w,x) \in \mathcal{B}_{1,w,x}$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$ satisfying (1)*

Theorem 2.1 provides a necessary and sufficient characterization of expected utility maximization behavior that only involves the data and the bounds on the choice-dependent potential outcome probabilities. Importantly, the characterization no longer depends on the decision maker's private information, which allows researchers to statistically test whether these inequalities are satisfied. It builds on the "no-improving action switches" inequalities in Caplin and Martin (2015), which characterize Bayesian expected utility maximization behavior in state-dependent stochastic choice

environments. By identifying states of the world (i.e., mappings from choices to an abstract prize space) with the unobserved potential outcomes and characteristics, an abstract prize space with the decision maker's utility payoffs over choices, potential outcomes, and characteristics, and incorporating the missing data problem, I obtain a tractable characterization of expected utility maximization in a wide range of empirical settings. By applying these revealed preference inequalities in decision problems of interest, researchers can derive interpretable conditions on behavior. In the next section, I use Theorem 2.1 to analyze the large class of screening decisions.

The proof of sufficiency for Theorem 2.1 shows that if the revealed preference inequalities (1) are satisfied, then a joint distribution $(W, X, V, C, \vec{Y}) \sim Q$ can be constructed under the expected utility maximization model that satisfies Data Consistency (Definition 3).[21] I construct a likelihood function for the private information $Q_V(\cdot \mid \vec{y}, w, x)$ such that if the decision maker finds it optimal to select choice $C = 0, C = 1$, then the decision maker's posterior beliefs $Q_{\vec{Y}}(\cdot \mid w, x, v)$ equal the choice-dependent outcome probabilities $\widetilde{P}_{\vec{Y}}(\cdot \mid 0, W, X)$, $P_{\vec{Y}}(\cdot \mid 1, w, x)$ respectively. By this construction, these choice-dependent potential outcome probabilities are therefore Bayes-plausible posterior beliefs with respect to some conditional distribution for the potential outcomes given the characteristics in the identified set $\mathcal{H}_P(P_{\vec{Y}}(\cdot \mid w, x); \mathcal{B}_{w,x})$. The revealed preference inequalities (1) imply that this construction satisfies Expected Utility Maximization, and additional work remains to show that it also satisfies Data Consistency (Definition 3). In this sense, Theorem 2.1 establishes that the choice-dependent potential outcome probabilities summarize the decision maker's posterior beliefs under any distribution of private information. The researcher's assumptions about the missing data therefore have a behavioral interpretation as restrictions on the possible informativeness of the decision maker's private information. Supplement G.1 shows that the researchers' bounds on the missing data (Assumption 2.1) restrict the average informativeness of the decision maker's private information in a screening decision.

3 Testing Expected Utility Maximization in Screening Decisions

In this section, I use Theorem 2.1 to characterize the testable implications of expected utility maximization in screening decisions with a binary outcome, such as pretrial release and medical testing, under various assumptions on the decision maker's preferences and the missing data problem. Testing these restrictions is equivalent to testing many moment inequalities, and I discuss how supervised machine learning based methods may be used to reduce the dimension on this testing problem.

[21] While its proof is constructive, Theorem 2.1 can also be established through Bergemannn and Morris (2016)'s equivalence result between the set of Bayesian Nash Equilibrium and the set of Bayes Correlated Equilibrium in incomplete information games, where the potential outcome vector $\vec{Y}$ and the characteristics (W, X) are the state, the initial information structure is the null information structure, the private information V is the augmenting signal structure, and applying Data Consistency (Definition 3) on the equilibrium conditions.

3.1 Characterization in Screening Decisions

In a screening decision, the potential outcome under the decision maker's choice $C = 0$ satisfies $Y_0 \equiv 0$ and $Y_1 := Y^*$ is a latent outcome that is revealed whenever the decision maker selected choice $C = 1$. For exposition, I further assume that the latent outcome is binary $\mathcal{Y} = \{0, 1\}$ as in the motivating applications of pretrial release and medical testing. Appendix C.1 extends these results to treatment decisions with a scalar, multi-valued outcome for particular classes of utility functions.

Focusing on a screening decision with a binary outcome simplifies the setting in Section 2. The bounds on the choice-dependent outcome probabilities given choice $C = 0$ are an interval for the conditional probability of $Y^* = 1$ given choice $C = 0$ with $\mathcal{B}_{0,w,x} = [\underline{P}_{Y^*}(1 \mid 0, w, x), \overline{P}_{Y^*}(1 \mid 0, w, x)]$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$. The bounds given choice $C = 1$ are the point identified conditional probability of $Y^* = 1$ given choice $C = 1$ with $\mathcal{B}_{1,w,x} = \{P_{Y^*}(1 \mid 1, w, x)\}$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$. Finally, it is without loss of generality to normalize two entries of the utility function $U(c, y^*; w)$, and so I normalize $U(0, 1; w) = 0, U(1, 0; w) = 0$ for all $w \in \mathcal{W}$.[22]

I derive conditions under which the decision maker's choices are consistent with expected utility maximization at strict preferences, meaning the decision maker strictly prefers a unique choice at each latent outcome under the expected utility maximization model. This rules out trivial cases such as complete indifference.

Definition 6 (Strict Preferences)**.** The utility functions $U \in \mathcal{U}$ satisfy *strict preferences* if $U(0, 0; w) < 0$ and $U(1, 1; w) < 0$ for all $w \in \mathcal{W}$.

In the pretrial release example, focusing on strict preference utility functions means that the researcher is willing to assume it is always costly for the judge to either detain a defendant ($C = 0$) that would not commit pretrial misconduct ($Y^* = 0$) or release a defendant ($C = 1$) that would commit pretrial misconduct ($Y^* = 1$).

By applying Theorem 2.1, I characterize the conditions under which the decision maker's choices in a screening decision with a binary outcome are consistent with expected utility maximization at some strict preference utility function and private information. For each $w \in \mathcal{W}$, define $\mathcal{X}^1(w) := \{x \in \mathcal{X} \colon \pi_1(w, x) > 0\}$ and $\mathcal{X}^0(w) := \{x \in \mathcal{X} \colon \pi_0(w, x) > 0\}$.

Theorem 3.1. *Consider a screening decision with a binary outcome. Assume $P_{Y^*}(1 \mid 1, w, x) < 1$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $\pi_1(w, x) > 0$. The decision maker's choices are consistent with expected utility maximization at some strict preference utility function if and only if, for all $w \in \mathcal{W}$,*

$$\max_{x \in \mathcal{X}^1(w)} P_{Y^*}(1 \mid 1, w, x) \leq \min_{x \in \mathcal{X}^0(w)} \overline{P}_{Y^*}(1 \mid 0, w, x).$$

[22] In some settings, it may be more natural to instead normalize $U(0, 0; w) = 0, U(1, 1; w) = 0$.

Otherwise, $\mathcal{H}_P(U;\mathcal{B}) = \emptyset$, and the decision maker is making detectable prediction mistakes based on the observable characteristics.

Corollary 3.1. *The identified set of strict preference utility functions $\mathcal{H}_P(U;\mathcal{B})$ is equal to the set of all utility functions satisfying, for all $w \in \mathcal{W}$, $U(0,0;w) < 0$, $U(1,1;w) < 0$ and*

$$\max_{x \in \mathcal{X}^1(w)} P_{Y^*}(1 \mid 1, w, x) \leq \frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \leq \min_{x \in \mathcal{X}^0(w)} \overline{P}_{Y^*}(1 \mid 0, w, x).$$

In a screening decision with a binary outcome, expected utility maximization at strict preferences requires the decision maker to make choices according to an incomplete threshold rule based on their posterior beliefs given the characteristics and private information. The threshold only depends on the characteristics W, and it is incomplete since it takes no stand on how possible indifferences are resolved. As mentioned, Theorem 2.1 establishes that the choice-dependent outcome probabilities summarize all possible posterior beliefs under the expected utility maximization model. Applying an incomplete threshold rule to posterior beliefs under the expected utility maximization model is therefore observationally equivalent to applying a threshold rule to the choice-dependent outcome probabilities. Theorem 3.1 formalizes this argument, and shows that the decision maker's choices are consistent with expected utility maximization at some strict preference utility function if and only if there exists some value of the unobservable choice-dependent outcome probabilities that are consistent with the researcher's bounds (Assumption 2.1) and would reproduce the decision maker's observed choices under such a threshold rule. The threshold $\frac{U(0,0;w)}{U(0,0;w)+U(1,1;w)}$ itself summarizes the relative costs of "ex-post errors" – that is, the cost of choosing $C = 0$ for an individual with outcome $Y^* = 0$ relative to the cost of choosing $C = 1$ for an individual with outcome $Y^* = 1$.

If the conditions in Theorem 3.1 are violated, there exists no strict preference utility function nor private information such that the decision maker's choices are consistent with expected utility maximization at accurate beliefs. By examining cases in which these conditions are satisfied, I next characterize conditions under which we cannot identify whether the decision maker makes detectable prediction mistakes based on the characteristics.

First, Theorem 3.1 highlights the necessity of placing an exclusion restriction on which observable characteristics directly affect the decision maker's utility function. If all observable characteristics directly affect the decision maker's utility function (i.e., $\mathcal{X} = \emptyset$), then the decision maker's choices are consistent with expected utility maximization whenever the researcher assumes the decision maker observes useful private information.

Corollary 3.2. *Under the same conditions as Theorem 3.1, suppose $\mathcal{X} = \emptyset$ and all observable characteristics therefore directly affect the decision maker's utility function. If $P_{Y^*}(1 \mid 1, w) \leq$*

$\overline{P}_{Y^*}(1 \mid 0, w)$ *for all* $w \in \mathcal{W}$*, then the decision maker's choices are consistent with expected utility maximization at some strict preference utility function.*

This negative result arises because a characteristic-dependent threshold can always be constructed that rationalizes the decision maker's observed choices if the researcher's assumptions allow the probability of $Y^* = 1$ given $C = 0$ to be at least as large as the observed probability of $Y^* = 1$ given $C = 1$ for all characteristics. In this case, the decision maker's observed choices are consistent with expected utility maximization at a strict preference utility function that varies richly across the characteristics $w \in \mathcal{W}$. If the researcher suspects that the decision maker observes useful private information, then imposing an exclusion restriction that some observable characteristics do not directly enter into the utility function is necessary for identifying prediction mistakes.

Unfortunately, imposing such an exclusion restriction is not alone sufficient to restore the testability of expected utility maximization. The researcher must still address the missing data problem.

Corollary 3.3. *Under the same conditions as Theorem 3.1, if* $\overline{P}_{Y^*}(1 \mid 0, w, x) = 1$ *for all* $(w, x) \in \mathcal{W} \times \mathcal{X}$*, then the decision maker's choices are consistent with expected utility maximization at some strict preferences.*

Without informative bounds on the unobservable choice-dependent 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 (i.e., $\overline{P}_{Y^*}(1 \mid 0, w, x) = 1$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$).

Corollaries 3.2-3.3 highlight that testing expected utility maximization, and therefore detectable prediction mistakes, requires both behavioral assumptions on which characteristics may directly affect the decision maker's preferences and econometric assumptions that generate informative bounds on the unobservable choice-dependent outcome probabilities. Under such assumptions, Theorem 3.1 provides interpretable conditions to test for detectable prediction mistakes. At any fixed $w \in \mathcal{W}$, does there exist some $x \in \mathcal{X}$ such that the largest possible probability of $Y^* = 1$ given $C = 0$ is strictly lower than the observed probability of outcome $Y^* = 1$ given $C = 1$ at some other $x' \in \mathcal{X}$? If so, then the decision maker cannot be maximizing expected utility as the decision maker could do strictly better by raising their probability of selecting choice $C = 0$ at x' and lowering their probability of selecting choice $C = 1$ at x. Theorem 3.1 shows that these "misranking" arguments are necessary and sufficient to test the joint null hypothesis that the decision maker's choices are consistent with expected utility maximization at accurate beliefs and their preferences satisfy the conjectured exclusion restriction.

Example (Pretrial Release)**.** Theorem 3.1 requires that, holding fixed the defendant characteristics that directly affect the judge's preferences, all detained defendant must have a higher worst-case

probability of committing pretrial misconduct than any released defendant. Suppose the researcher assumes that the judge may engage in taste-based discrimination based on the defendant race W, but the judge's preferences are unaffected by remaining defendant characteristics X. To test for detectable prediction mistakes, the researcher must check, among defendants of the same race, whether there exists some group of released defendants with a higher pretrial misconduct than the worst-case pretrial misconduct rate of some group of detained defendants among defendants. In this case, the judge must be misranking the probability of pretrial misconduct, and their choices are inconsistent with expected utility maximization at strict preferences that only depend on defendant race. ▲

Remark 3.1 (Connection to Marginal Outcome Tests and Inframarginality)**.** In a screening decision with a binary outcome, my analysis complements recent results in Canay, Mogstad and Mountjoy (2020), Gelbach (2021), and Hull (2021), which use an extended Roy model to explore the validity of marginal outcome tests for taste-based discrimination. This literature exploits the continuity of private information to derive testable implications of extended Roy selection in terms of underlying marginal treatment effect functions, which requires that the researcher identify the conditional expectation of the outcome at each possible "marginal" decision. In contrast, Theorem 3.1 involves only "inframarginal" outcomes. Common sources of quasi-experimental variation lead to bounds on inframarginal outcomes without additional assumptions such as monotonicity or functional form restrictions needed to estimate marginal treatment effect curves.

Remark 3.2 (Approximate Expected Utility Maximization)**.** This identification analysis assumes that the decision maker fully maximizes expected utility. The decision maker, however, may face cognitive constraints that prevent them from doing so completely. Appendix C.2 considers a generalization in which the decision maker "approximately" maximizes expected utility, meaning that the decision maker's choices must only be within $\epsilon_w \geq 0$ of optimal. The decision maker is boundedly rational in this sense, and selects a choice to "satisfice" expected utility (Simon, 1955, 1956). Expected utility maximization (Definition 3) is the special case with $\epsilon_w = 0$, and any behavior is rationalizable as approximately maximizing expected utility for ϵ_w sufficiently large. The smallest, rationalizing parameter $\epsilon_w \geq 0$ summarizes how far from optimal are the decision maker's observed choices, and I show that researchers can conduct inference on this object.

3.2 Constructing Bounds on the Missing Data with an Instrument

Suppose there is a randomly assigned instrument that generates variation in the decision maker's choice probabilities in a screening decision. Such instruments commonly arise, for example, through the random assignment of decision makers to individuals. Judges are randomly assigned to defendants in pretrial release (Kling, 2006; Dobbie, Goldin and Yang, 2018; Arnold, Dobbie

and Yang, 2018; Kleinberg et al., 2018*a*; Arnold, Dobbie and Hull, 2020*b*) and doctors may be randomly assigned to patients in medical testing (Abaluck et al., 2016; Chan, Gentzkow and Yu, 2020).[23]

Assumption 3.1 (Random Instrument)**.** Let $Z \in \mathcal{Z}$ be a finite support instrument, and the joint distribution $(W, X, Z, C, Y^*) \sim P$ satisfies $(W, X, Y^*) \perp\!\!\!\perp Z$ and $P(W = w, X = x, Z = z) > 0$ for all $(w, x, z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Z}$.

The researcher observes the joint distribution $(W, X, Z, C, Y) \sim P$, where $Y = C \cdot Y^*$ as before. The goal is to construct bounds on the unobservable choice-dependent outcome probabilities at each observable characteristic $(w, x) \in \mathcal{W} \times \mathcal{X}$ and at a particular value of the instrument $z \in \mathcal{Z}$. In the case where decision makers are randomly assigned to cases, this corresponds to constructing bounds for a particular decision maker, which can then be used to test whether that decision maker is making detectable prediction mistakes using the identification results in the previous section.

Under Assumption 3.1, the unobservable choice-dependent outcome probabilities are partially identified, denoting their sharp identified sets as $\mathcal{H}_P(P_{Y^*}(\cdot \mid 0, w, x, z))$.

Proposition 3.1. *Suppose Assumption 3.1 holds and consider a screening decision with a binary outcome. Then, for any* $(w, x, z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Z}$ *with* $\pi_0(w, x, z) > 0$*,* $\mathcal{H}_P(P_{Y^*}(\cdot \mid 0, w, x, z)) = [\underline{P}_{Y^*}(1 \mid 0, w, x, z), \overline{P}_{Y^*}(1 \mid 0, w, x, z)]$*, where*

$$\underline{P}_{Y^*}(1 \mid 0, w, x, z) = \max\left\{\frac{\underline{P}_{Y^*}(1 \mid w, x) - P_{C,Y^*}(1, 1 \mid w, x, z)}{\pi_0(w, x, z)}, 0\right\},$$

$$\overline{P}_{Y^*}(1 \mid 0, w, x, z) = \min\left\{\frac{\overline{P}_{Y^*}(1 \mid w, x) - P_{C,Y^*}(1, 1 \mid w, x, z)}{\pi_0(w, x, z)}, 1\right\},$$

and $\underline{P}_{Y^*}(1 \mid w, x) = \max_{\tilde{z} \in \mathcal{Z}}\{P_{C,Y^*}(1, 1 \mid w, x, \tilde{z})\}$*,* $\overline{P}_{Y^*}(1 \mid w, x) = \min_{\tilde{z} \in \mathcal{Z}}\{\pi_0(w, x, \tilde{z}) + P_{C,Y^*}(1, 1 \mid w, x, \tilde{z})\}$.

The bounds in Proposition 3.1 follow from worst-case bounds on $P_{Y^*}(1 \mid w, x)$ (e.g., Manski, 1989, 1994) and point identification of $P_{C,Y^*}(1, 1 \mid w, x, z), \pi_0(w, x, z)$.[24,25] Therefore, for a fixed

[23] There are other examples of instruments in empirical research. For example, Mullainathan and Obermeyer (2020) argue that doctors are less likely to conduct stress tests for a heart attack on Fridays and Saturdays due to weekend staffing constraints, even though patients that arrive on these days are no less risky. The introduction of or changes to recommended guidelines may also affect decision makers' choices (Albright, 2019; Abaluck et al., 2020).

[24] An active literature focuses on the concern that the monotonicity assumption in Imbens and Angrist (1994) may be violated in settings where decision makers are randomly assigned to individuals. de Chaisemartin (2017) and Frandsen, Lefgren and Leslie (2019) develop weaker notions of monotonicity for these settings. Proposition 3.1 imposes no form of monotonicity (or its relaxations) in deriving bounds.

[25] Lakkaraju et al. (2017) and Kleinberg et al. (2018*a*) use the random assignment of decision makers to evaluate a statistical decision rule $\tilde{C}$ by imputing its true positive rate $P(Y^* = 1 \mid \tilde{C} = 1)$. In contrast, Proposition 3.1 constructs bounds on a decision maker's conditional choice-dependent outcome probabilities, $P_{Y^*}(1 \mid 0, w, x)$.

value $z \in \mathcal{Z}$, the researcher may apply the identification results derived in Section 3.1 by defining $\mathcal{B}_{0,w,x} = \mathcal{H}_P(P_{Y^*}(\cdot \mid 0, w, x, z))$ under Assumption 3.1.

Appendix D.1 extends these bounds to allow for the instrument to be quasi-randomly assigned conditional on some additional characteristics, which will be used in the empirical application to pretrial release decisions in New York City. Appendix D.2 extends these bounds to treatment decisions.

Under the expected utility maximization model, Assumption 3.1 requires that the decision maker's beliefs about the latent outcome given the observable characteristics do not depend on the instrument.

Proposition 3.2. *Suppose Assumption 3.1 holds. If the decision maker's choices are consistent with expected utility maximization at some utility function U and joint distribution $(W, X, Z, V, C, Y^*) \sim Q$, then $Y^* \perp\!\!\!\perp Z \mid W, X$ under Q.*

This is a consequence of Data Consistency in Definition 3. Requiring that the decision maker's beliefs about the outcome given the characteristics be accurate imposes that the instrument cannot affect their beliefs about the outcome given the observable characteristics if it is randomly assigned. Aside from this restriction, Assumption 3.1 places no further behavioral restrictions on the expected utility maximization model. Consider the pretrial release setting in which the instrument arises through the random assignment of judges, meaning $Z \in \mathcal{Z}$ refers to a judge identifier. Proposition 3.2 implies that if all judges make choices as-if they are maximizing expected utility at accurate beliefs given defendant characteristics, then all judges must have the same beliefs about the probability of pretrial misconduct given defendant characteristics. Judges may still differ from one another in their preferences and private information.

Remark 3.3 (Other empirical strategies for constructing bounds)**.** Supplement G discusses two additional empirical strategies for constructing bounds on the unobservable choice-dependent outcome probabilities. First, researchers may use the observable choice-dependent outcome probabilities to directly bound the unobservable choice-dependent outcome probabilities. Second, the researcher may use a "proxy outcome," which does not suffer the missing data problem and is correlated with the outcome. For example, researchers often use future health outcomes as a proxy for whether patients had a particular underlying condition at the time of the medical testing or diagnostic decision (Mullainathan and Obermeyer, 2020; Chan, Gentzkow and Yu, 2020).

3.3 Reduction to Moment Inequalities

Testing whether the decision maker's choices in a screening decision with a binary outcome are consistent with expected utility maximization at strict preferences (Theorem 3.1) reduces to testing many moment inequalities.

Proposition 3.3. *Consider a screening decision with a binary outcome. Suppose Assumption 3.1 holds, and* $0 < \pi_c(w,x,z) < 1$ *for all* $(w,x,z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Z}$*. The decision maker's choices at* $z \in \mathcal{Z}$ *are consistent with expected utility maximization at some strict preference utility function if and only if for all* $w \in \mathcal{W}$*, pairs* $x, \tilde{x} \in \mathcal{X}$ *and* $\tilde{z} \in \mathcal{Z}$

$$P_{Y^*}(1 \mid 1,w,x,z) - \overline{P}_{Y^*,\tilde{z}}(1 \mid 0,w,\tilde{x},z) \leq 0,$$

where $\overline{P}_{Y^*,\tilde{z}}(1 \mid 0,w,x,z) = \frac{\pi_0(w,x,\tilde{z}) + P_{C,Y^*}(1,1|w,x,\tilde{z})}{\pi_0(w,x,z)} - \frac{P_{C,Y^*}(1,1|w,x,z)}{\pi_0(w,x,z)}$.

The number of moment inequalities is equal to $d_w \cdot d_x^2 \cdot (d_z - 1)$, and grows with the number of support points of the characteristics and instruments. In empirical applications, this will be quite large since the characteristics of individuals are extremely rich. Testing the revealed preference inequalities in a screening decision with a binary outcome directly over the underlying characteristics may require using moment inequality procedures that are valid in high-dimensional settings such as Chernozhukov, Chetverikov and Kato (2019) and Bai, Santos and Shaikh (2021).

The number of moment inequalities may be reduced by testing implied revealed preference inequalities over any partition of the excluded characteristics. For each $w \in \mathcal{W}$, define $D_w \colon \mathcal{X} \to \{1, \ldots, N_d\}$ to be some function that partitions the support of the excluded characteristics $x \in \mathcal{X}$ into level sets $\{x \colon D_w(x) = d\}$. By iterated expectations, if the decision maker's choices are consistent with expected utility maximization at some strict preference utility function, then their choices must satisfy *implied* revealed preference inequalities. Define $P_{Y^*}(y^* \mid c,w,d) := P_{Y^*}(Y^* = y^* \mid C = c, W = w, D_w(X) = d)$ and $\pi_c(w,d) := P(C = c \mid W = w, D_w(X) = d)$.

Corollary 3.4. *Consider a screening decision with a binary outcome, and assume* $P_{Y^*}(1 \mid 1,w,x) < 1$ *for all* $(w,x) \in \mathcal{W} \times \mathcal{X}$ *with* $\pi_1(w,x) > 0$*. If the decision maker's choices are consistent with expected utility maximization at some strict preference utility function, then for all* $w \in \mathcal{W}$

$$\max_{d \in \mathcal{D}^1(w)} P_{Y^*}(1 \mid 1,w,d) \leq \min_{x \in \mathcal{D}^0(w)} \overline{P}_{Y^*}(1 \mid 0,w,d),$$

where $\mathcal{D}^1(w) := \{d \colon \pi_1(w,d) > 0\}$ *and* $\mathcal{D}^0(w) := \{d \colon \pi_0(w,d) > 0\}$.

In practice, this can drastically reduce the number of moment inequalities that must be tested. If $N_d \ll d_x$, researchers may instead test implied revealed preference inequalities in Corollary 3.4 using procedures that are valid in low-dimensional settings, which is a mature literature in econometrics. See, for example, the reviews in Canay and Shaikh (2017), Ho and Rosen (2017) and Molinari (2020).

A natural choice is to construct the partitioning functions $D_w(\cdot)$ using supervised machine learning methods that predict the outcome on a set of held-out decisions. Suppose the researcher

estimates the prediction function $\hat{f}\colon \mathcal{W} \times \mathcal{X} \to [0,1]$ on a set of held-out decisions. Given the estimated prediction function, the researcher may define $D_w(x)$ by binning the characteristics X into percentiles of predicted risk within each value $w \in \mathcal{W}$. The resulting implied revealed preference inequalities search for misrankings in the decision maker's choices across percentiles of predicted risk. Alternatively, 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 medical testing or treatment decisions, commonly used risk assessments summarize observable patient characteristics into an integer-valued risk score (e.g., Obermeyer and Emanuel, 2016; Lakkaraju and Rudin, 2017). The researcher may therefore define the partition $D_w(x)$ to be level sets associated with this existing risk score.

Corollary 3.4 clarifies how precisely supervised machine learning methods can be used to test for prediction mistakes in a large class of "prediction policy problems" (Kleinberg et al., 2015). In existing empirical work, researchers directly compare the recommended choices of an estimated decision rule or prediction function against the choices of a human decision maker. See, for example, Meehl (1954), Dawes, Faust and Meehl (1989), Grove et al. (2000) in psychology, Kleinberg et al. (2018*a*), Ribers and Ullrich (2019), Mullainathan and Obermeyer (2020) in economics, and Chouldechova et al. (2018), Jung et al. (2020*a*) in computer science. Such comparisons rely on strong assumptions that restrict preferences to be constant across both decisions and decision makers (Kleinberg et al., 2018*a*; Mullainathan and Obermeyer, 2020) or that observed choices were as-good-as randomly assigned given the characteristics (Lakkaraju and Rudin, 2017; Chouldechova et al., 2018; Jung et al., 2020*a*). In contrast, these results show that such prediction models should instead be used to construct partitions of the characteristics that are assumed to not directly affect preferences. Given such a partition, checking whether the decision maker's choices satisfy the implied revealed preference inequalities in Corollary 3.4 is a valid test for whether the decision maker's choices maximize expected utility at some strict preferences that may arbitrarily vary across directly payoff-relevant characteristics and any private information. This provides, to the best of my knowledge, the first microfounded procedure for using such estimated prediction functions to formally test whether the decision maker's choices are consistent with expected utility maximization at accurate beliefs.

Appendix D.2 extends these results to treatment decisions, showing that testing whether the decision maker's choices are consistent with expected utility maximization in a treatment decision reduces to testing a system of moment inequalities with nuisance parameters that enter linearly.

4 Bounding Prediction Mistakes based on Characteristics

So far, I have shown that researchers can test whether a decision maker's choices are consistent with expected utility maximization at accurate beliefs about the outcome, and therefore whether the decision maker is making detectable prediction mistakes. By modifying the expected utility maximization model and the revealed preference inequalities, researchers can further investigate the ways in which the decision maker's predictions are systematically biased.

4.1 Expected Utility Maximization at Inaccurate Beliefs

The definition of expected utility maximization (Definition 3) implied that the decision maker acted as-if their implied beliefs about the outcome given the characteristics were accurate (Lemma 2.1). As a result, the revealed preference inequalities may be violated if the decision maker acted as-if they maximized expected utility based on *inaccurate* beliefs, meaning that their implied beliefs do not lie in the identified set for the distribution of the outcome given the characteristics $\mathcal{H}_P(P_{\vec{Y}}(\cdot \mid w,x);\mathcal{B}_{w,x})$. This is a common behavioral hypothesis in empirical applications. For example, 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 testing and treatment decisions.[26]

To investigate whether the decision maker's choices may maximize expected utility at inaccurate beliefs, I modify "Data Consistency" in Definition 3.

Definition 7. 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 $(W,X,V,C,\vec{Y}) \sim Q$ satisfying (i) Expected Utility Maximization, (ii) Information Set, and

iii. Data Consistency with Inaccurate Beliefs: For all $(w,x) \in \mathcal{W} \times \mathcal{X}$, there exists $\widetilde{P}_{\vec{Y}}(\cdot \mid 0,w,x) \in \mathcal{B}_{0,w,x}$ and $\widetilde{P}_{\vec{Y}}(\cdot \mid 1,w,x) \in \mathcal{B}_{1,w,x}$ such that for all $\vec{y} \in \mathcal{Y}^2$ and $c \in \{0,1\}$

$$Q_C(c \mid \vec{y},w,x)\widetilde{P}_{\vec{Y}}(\vec{y} \mid w,x)Q(w,x) = \widetilde{P}_{\vec{Y}}(\vec{y} \mid c,w,x)P(c,w,x),$$

where $\widetilde{P}_{\vec{Y}}(\vec{y} \mid w,x) = \widetilde{P}_{\vec{Y}}(\vec{y} \mid 0,w,x)\pi_0(w,x) + \widetilde{P}_{\vec{Y}}(\vec{y} \mid 1,w,x)\pi_1(w,x)$.

Definition 7 only requires that the joint distribution $(W,X,V,C,\vec{Y}) \sim Q$ under the expected utility maximization model matches the observable joint distribution P if the decision maker's model-implied beliefs given the characteristics, $Q_{\vec{Y}}(\cdot \mid w,x)$, are replaced with some marginal

[26] 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).

distribution of the outcome given the characteristics that lies in the identified set, $\widetilde{P}_{\vec{Y}}(\cdot \mid w,x) \in \mathcal{H}_P(P_{\vec{Y}}(\cdot \mid w,x); \mathcal{B}_{w,x})$. This imposes that the decision maker is acting as-if they correctly specified the likelihood of their private information $V \mid \vec{Y}, W, X$, their prediction mistakes only arise from their inaccurate prior beliefs about the outcome given the characteristics. Definition 7 places no restrictions on the decision maker's implied prior beliefs $Q_{\vec{Y}}(\cdot \mid w,x)$, so behavior that is consistent with expected utility maximization at inaccurate beliefs could arise from several alternative behavioral hypotheses. The decision maker's implied prior beliefs may, for instance, be 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).

The next result characterizes whether the decision maker's observed choices are consistent with expected utility maximization at inaccurate beliefs in terms of modified revealed preference inequalities.

Theorem 4.1. *Assume $\widetilde{P}_{\vec{Y}}(\cdot \mid w,x) > 0$ for all $\widetilde{P}_{\vec{Y}}(\cdot \mid w,x) \in \mathcal{H}_P(P_{\vec{Y}}(\cdot \mid w,x); \mathcal{B}_{w,x})$ and all $(w,x) \in \mathcal{W} \times \mathcal{X}$. The decision maker's choices are consistent with expected utility maximization at inaccurate beliefs if and only if there exists a utility function $U \in \mathcal{U}$, $\widetilde{P}_{\vec{Y}}(\cdot \mid 0,w,x) \in \mathcal{B}_{0,w,x}$ and $\widetilde{P}_{\vec{Y}}(\cdot \mid 1,w,x) \in \mathcal{B}_{1,w,x}$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$, and non-negative weights $\omega(\vec{y}; w,x)$ satisfying*

i. For all $c \in \{0,1\}$, $(w,x) \in \mathcal{W} \times \mathcal{X}$ with $\pi_c(w,x) > 0$, $c' \neq c$

$$\mathbb{E}_{\widetilde{P}}\left[\omega(\vec{Y}; W, X) U(c, \vec{Y}; W) \mid C = c, W = w, X = x\right] \geq$$

$$\mathbb{E}_{\widetilde{P}}\left[\omega(\vec{Y}; W, X) U(c', \vec{Y}; W) \mid C = c, W = w, X = x\right]$$

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

where $\mathbb{E}_{\widetilde{P}}[\cdot]$ is the expectation under the joint distribution under $(W, X, C, Y^*) \sim \widetilde{P}$ defined as $\widetilde{P}(w,x,c,\vec{y}) = \widetilde{P}_{Y^*}(\vec{y} \mid c,w,x) P(c,w,x)$.

These modified inequalities ask whether the decision maker's observed choices satisfy revealed preference inequalities at a reweighed utility function, where the weights $\omega(\vec{y}; w,x)$ are the likelihood ratio of the decision maker's implied prior beliefs relative to some conditional distribution of the potential outcomes given the characteristics in the identified set. Since the decision maker's prediction mistakes only arise from misspecification of prior beliefs $Q_{\vec{Y}}(\vec{Y} \mid w,x)$, her posterior beliefs under the model are proportional to the likelihood ratio between her prior beliefs and the underlying potential outcome distribution. These likelihood ratio weights can be interpreted as a modification of the decision maker's utility function since expected utility is linear in beliefs and utility. This suggests that expected utility maximization at inaccurate beliefs is equivalent to expected utility maximization at accurate beliefs and preferences that are summarized by this

reweighed utility function. Theorem 4.1 formalizes this intuition and shows that it completely characterizes whether the decision maker's choices are consistent with expected utility maximization at inaccurate beliefs.

4.2 Bounding Prediction Mistakes in Screening Decisions

Theorem 4.1 implies that researchers can bound the extent to which the decision maker's prior beliefs given the characteristics overreacts or underreacts to variation in the characteristics in a screening decision with a binary outcome.[27] This enables researchers to report interpretable bounds on the extent to which the decision maker's predictions are systematically biased.

As a first step, Theorem 4.1 implies bounds on the decision maker's reweighted utility function $\omega(y^*; w, x)U(c, y^*; w)$ in a screening decision with a binary outcome.

Theorem 4.2. *Consider a binary screening decision. Assume* $0 < P_{Y^*}(1 \mid w, x) < 1$ *for all* $P_{Y^*}(\cdot \mid w, x) \in \mathcal{H}_P(P_{Y^*}(\cdot \mid w, x); \mathcal{B}_{w,x})$ *and* $0 < \pi_1(w, x) < 1$ *for all* $(w, x) \in \mathcal{W} \times \mathcal{X}$*. Suppose the decision maker's choices are consistent with expected utility maximization at inaccurate beliefs and some strict preference utility function. Then, there exists non-negative weights* $\omega(y^*; w, x) \geq 0$ *satisfying for all* $(w, x) \in \mathcal{W} \times \mathcal{X}$

$$P_{Y^*}(1 \mid 1, w, x) \leq \frac{\omega(0; w, x)U(0, 0; w)}{\omega(0; w, x)U(0, 0; w) + \omega(1; w, x)U(1, 1; w)} \leq \overline{P}_{Y^*}(1 \mid 0, w, x), \tag{2}$$

where $\omega(y^*; w, x) = Q_{Y^*}(y^* \mid w, x,)/\widetilde{P}_{Y^*}(y^* \mid w, x)$ *and* $Q_{Y^*}(y^* \mid w, x)$*,* $\widetilde{P}_{Y^*}(y^* \mid w, x)$ *are defined in Definition 7.*

That is, in a screening decision with a binary outcome, expected utility maximization at inaccurate beliefs is observationally equivalent to an incomplete threshold rule based on the choice-dependent outcome probabilities, where the threshold now depends on the decision maker's reweighed utility function. This result may be exploited to derive an identified set on the extent to which the decision maker overreacts or underreacts to variation in the characteristics. Define $\delta(w, x) := \frac{Q_{Y^*}(1|w,x)/Q_{Y^*}(0|w,x)}{\widetilde{P}_{Y^*}(1|w,x)/\widetilde{P}_{Y^*}(0|w,x)}$ to be the relative odds ratio of the unknown outcome under the decision maker's implied beliefs relative to the true conditional distribution, and $\tau(w, x) := \frac{\omega(0;w,x)U(0,0;w)}{\omega(0;w,x)U(0,0;w)+\omega(1;w,x)U(1,1;w)}$ to be the decision maker's reweighed utility threshold. If the reweighed utility threshold were known, then the decision maker's implied prediction mistake could be backed out.

[27] These results generalized directly to multi-valued outcomes over the class of binary-valued utility functions. That is, for some known $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$, define $\tilde{Y}^* = 1\{Y^* \in \tilde{\mathcal{Y}}\}$. The class of two-valued utility functions take the form $u(c, \vec{y}; w) := u(c, \tilde{y}; w)$ and satisfy strict preferences. For this class of utility function, the decision maker faces a screening decision with a binary outcome.

Corollary 4.1. *Under the same conditions as Theorem 4.2,*

$$\frac{(1-\tau(w,x))/\tau(w,x)}{(1-\tau(w,x'))/\tau(w,x')} = \frac{\delta(w,x)}{\delta(w,x')} \tag{3}$$

or any $w \in \mathcal{W}$, $x, x' \in \mathcal{X}$.

The ratio $\frac{\delta(w,x)}{\delta(w,x')}$ summarizes the extent to which the decision maker's implied beliefs about the outcome overreact or underreact to variation in the characteristics relative to the true conditional distribution. In particular, $\frac{\delta(w,x)}{\delta(w,x')}$ may be rewritten as the ratio of $\frac{Q_{Y^*}(1|w,x)/Q_{Y^*}(0|w,x)}{Q_{Y^*}(1|w,x')/Q_{Y^*}(0|w,x')}$ to $\frac{\widetilde{P}_{Y^*}(1|w,x)/\widetilde{P}_{Y^*}(0|w,x)}{\widetilde{P}_{Y^*}(1|w,x')/\widetilde{P}_{Y^*}(0|w,x')}$. The first term summarizes how the odds ratio of $Y^* = 1$ relative to $Y^* = 0$ varies across (w,x) and (w,x') under the decision maker's implied beliefs and the second term summarizes how the true odds ratio varies across the same values. If $\frac{\delta(w,x)}{\delta(w,x')}$ is less than one, then the decision maker's implied beliefs about the relative probability of $Y^* = 1$ versus $Y^* = 0$ react less to variation across the characteristics (w,x) and (w,x') than the true distribution. In this sense, their implied beliefs are *underreacting* across these characteristics. Analogously if $\frac{\delta(w,x)}{\delta(w,x')}$ is strictly greater than one, then the decision maker's implied beliefs about the relative probability of $Y^* = 1$ versus $Y^* = 0$ are *overreacting* across the characteristics (w,x) and (w,x'). Since Theorem 4.2 provides an identified set for the reweighted utility thresholds, an identified set for the implied prediction mistake $\frac{\delta(w,x)}{\delta(w,x')}$ can in turn be constructed by computing the ratio (3) for each pair $\tau(w,x), \tau(w,x')$ that satisfies (2).

Since the implied prediction mistake depends on the decision maker's perceived odds ratio across the characteristics $(w,x), (w,x')$, it could be consistent with multiple values of the decision maker's perceptions of baseline risk $Q_{Y^*}(1 \mid w,x), Q_{Y^*}(1 \mid w,x')$. As an example, suppose that $\widetilde{P}_{Y^*}(1 \mid w,x) = 4/5, \widetilde{P}_{Y^*}(1 \mid w,x') = 1/5$ and that the decision maker's perceptions of baseline risk are $Q_{Y^*}(1 \mid w,x) = 2/3, Q_{Y^*}(1 \mid w,x') = 1/3$. The true odds ratio of $Y^* = 1$ relative to $Y^* = 0$ at the characteristics are $\frac{\widetilde{P}_{Y^*}(1|w,x)}{\widetilde{P}_{Y^*}(0|w,x)} = 4$, $\frac{\widetilde{P}_{Y^*}(1|w,x')}{\widetilde{P}_{Y^*}(0|w,x')} = 1/4$. The decision maker's perceived odds ratio are $\frac{Q_{Y^*}(1|w,x)}{Q_{Y^*}(0|w,x)} = 2$, $\frac{Q_{Y^*}(1|w,x')}{Q_{Y^*}(0|w,x')} = 1/2$. In this case, the decision maker's implied prediction mistake $\frac{\delta(w,x)}{\delta(w,x')}$ equals $1/4$. If instead the decision maker's perceptions of baseline risk were $Q_{Y^*}(1 \mid w,x) = 3/4, Q_{Y^*}(1 \mid w,x') = 3/7$, then the decision maker's perceived odds ratios would equal $3, 3/4$ at characteristics $(w,x), (w,x')$ respectively. Even though the decision maker's perceptions of baseline risk are different, the implied prediction mistake again equals $1/4$. In both cases, the decision maker's perceptions across characteristics $(w,x), (w,x')$ underreact relative to the true variation in risk. The true odds ratio at (w,x) is 16 times larger than the true odds ratio at (w,x'), but the decision maker perceives it to be only 4 times larger. The implied prediction mistake summarizes how *relative* changes in the decision maker's beliefs across the characteristics compare to *relative* changes in the underlying distribution of outcomes.

Finally, these bounds on the implied prediction mistake are obtained only by specifying that

the decision maker's preferences satisfy the researcher's conjectured exclusion restriction and strict preferences. Under an exclusion restriction on preferences, variation in the decision maker's choices and outcomes across characteristics that do not directly affect utility must only arise due to variation in beliefs and variation in the underlying outcome probabilities. Using this intuition, these results show that examining how choice-dependent outcome probabilities vary across characteristics partially identifies a summary statistic of the decision maker's systematic prediction mistakes. Preference exclusion restrictions are therefore not only sufficient to test for systematic prediction mistakes, but are further sufficient to partially identify the extent to which variation in the decision maker's implied beliefs are biased.[28]

After applying the dimension reduction strategy in Section 3.3, the implied prediction mistake across values $D_w(X) = d, D_w(X) = d'$, now denoted by $\delta(w,d)/\delta(w,d')$, measures how the decision maker's implied beliefs of their own ex-post mistakes varies relative to the true probability of ex-post mistakes across values $D_w(X) = d, D_w(X) = d'$. This is still an informative summary of the decision maker's prediction mistakes. See Appendix C.3 for details.

5 Do Judges Make Prediction Mistakes?

As an empirical illustration, I apply this econometric framework to analyze the pretrial release decisions of judges in New York City, which is a leading example of a high-stakes screening decision.[29] I find that at least 20% of judges in New York City make detectable prediction mistakes in their pretrial release decisions. Under various exclusion restrictions on their preferences, their pretrial release decisions are inconsistent with expected utility maximization at accurate beliefs about failure to appear risk given defendant characteristics. These systematic prediction mistakes arise because judges fail to distinguish between predictably low risk and predictably high risk defendants. Rejections of expected utility maximization at accurate beliefs are driven primarily by release decisions on defendants at the tails of the predicted risk distribution.

5.1 Pretrial Release Decisions in New York City

I focus on 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 (2020*b*). The New York City pretrial system is an ideal setting to apply this econometric framework in three important ways. First, as discussed in Kleinberg et al. (2018*a*), the New York City pretrial system narrowly

[28]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)) and arises because the authors focus on settings in which there are no additional characteristics of decisions beyond those which directly affect utility.

[29]Because the data are sensitive and only available through an official data sharing agreement with the New York court system, I conducted this empirical analysis in conjunction with the University of Chicago Crime Lab (Rambachan and Ludwig, 2021).

asks judges to only consider failure to appear risk, not new criminal activity or public safety risk, in deciding whether to release a defendant. The latent outcome Y^* is therefore whether the defendant would fail to appear in court if released. Section 5.4.1 reports the robustness of my empirical findings to other choices of outcome Y^*. Second, the pretrial release system in New York City 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.

5.2 Data and Summary Statistics

I observe the universe of all arrests made in New York City between November 1, 2008 and November 1, 2013. This contains information on 1,460,462 cases, of which 758,027 cases were subject to a pretrial release decision.[30] I apply additional sample restrictions to construct the main estimation sample, which consists of 569,256 cases heard by 265 unique judges.[31,32] My empirical analysis tests whether each of the top 25 judges that heard the most cases make detectable prediction mistakes about failure to appear risk in their pretrial release decisions. These top 25 judges heard 243,118 cases in the main estimation sample. Each judge heard at least 5,000 cases in total (see Supplement Figure S1).

For each case, I observe demographic information about the defendant such as their race, gender, and age, information about the current charge filed against the defendant, the defendant's criminal record, and the defendant's 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.[33] If the defendant was released, I observe whether the defendant either failed to appear in court or was re-arrested for a new crime.

[30] To construct the set of cases that were subject to a pretrial release decision, I apply the sample restrictions used in Kleinberg et al. (2018*a*). This removes (i) removes desk appearance tickets, (ii) cases that were disposed at arraignment, and (iii) cases that were adjourned in contemplation of dismissal as well as duplicate cases.

[31] The following cases are further excluded: (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.

[32] Supplement Tables S4-S5 compare the main estimation sample to the universe of 758,027 cases that were subject to a pretrial release decision, broken out by the race of the defendant and by whether the defendant was released. Cases in the main estimation sample have more severe charges and a lower release rate than the universe of cases subject to a pretrial release decision.

[33] Judges in New York City decide whether to release defendants without conditions ("release on recognizance"), require that defendants post a chosen amount of bail, or deny bail altogether. Following prior empirical work on the pretrial release setting (e.g., Arnold, Dobbie and Yang, 2018; Kleinberg et al., 2018*a*; Arnold, Dobbie and Hull, 2020*b*), I collapse these choices into the binary decision of whether to release (either release on recognizance or set a bail among that the defendant pays) or detain (either set a bail amount that the defendant does not pay or deny bail altogether). I report the robustness of my findings to this choice in Section 5.4.1.

Supplement Table S1 provides descriptive statistics about the main estimation sample and the cases heard by the top 25 judges, broken out by the race of the defendant. 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. Supplement Table S2 reports the same descriptive statistics broken out by whether the defendant was released or detained, revealing that judges respond to defendant characteristics in their release decisions. Among defendants 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.

5.3 Empirical Implementation

I test whether the observed release decisions of judges in New York City maximize expected utility at accurate beliefs about failure to appear risk given defendant characteristics as well as some private information under various exclusion restriction on preferences. I test whether the revealed preference inequalities are satisfied assuming that (i) no defendant characteristics, (ii) the defendant's race, (iii) the defendant's race and age, and (iv) the defendant's race and whether the defendant was charged with a felony offense directly affects the judges' utility function. I discretize age into young and older defendants, where older defendants are those older than 25 years.

5.3.1 Constructing the Prediction Function

As a first step, I construct a partition of the excluded characteristics $X \in \mathcal{X}$ in order to reduce the number of moment inequalities as described in Section 3.3. I predict failure to appear $Y^* \in \{0, 1\}$ among defendants released by all other judges within each value of the payoff-relevant characteristics $W \in \mathcal{W}$, which are defined as either race-by-age cells or race-by-felony charge cells. The prediction function is an ensemble that averages the predictions of an elastic net model and a random forest.[34] 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 (see Supplement Figure S2).

[34] 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).

Both ensemble models achieve similar performance on released black and white defendants.

5.3.2 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.[35] 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, 2020*b*). I conduct these balance checks pooling together all defendants and separately within each payoff-relevant characteristic cell (defined by race-by-age cells or race-by-felony charge cells), reporting the coefficient estimates in Supplement Tables S6-S8. 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 unobservable failure to appear rate among defendants detained by each judge in the top 25. I group judges into quintiles of leniency based on the constructed leniency measure, and define the instrument $Z \in \mathcal{Z}$ to be the leniency quintile of the assigned judge. Applying the results in Appendix D.1, the bound on the failure to appear rate among defendants with $W = w, D_w(X) = d$ for a particular judge using leniency quintile $\tilde{z} \in \mathcal{Z}$ depends on quantities $\mathbb{E}[P(C = 1, Y^* = 1 \mid W = w, D_w(X) = d, Z = \tilde{z}, T) \mid W = w, D_w(X) = d]$ and $\mathbb{E}[P(C = 0 \mid W = w, D_w(X) = d, Z = \tilde{z}, T) \mid W = w, D_w(X) = d]$, 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 $P(C = 1, Y^* = 1 \mid W = w, D_w(x), Z = z, T = t)$ and $P(C = 0 \mid W = w, D_w(x), Z = z, T = t)$ as

$$1\{C = 1, Y^* = 1\} = \sum_{w,d,z} \beta^{c,y^*}_{w,d,z} 1\{W = w, D_w(X) = d, Z = z\} + \phi_t + \epsilon \tag{4}$$

$$1\{C = 0\} = \sum_{w,d,z} \beta^{c}_{w,d,z} 1\{W = w, D_w(X) = d, Z = z\} + \phi_t + \nu, \tag{5}$$

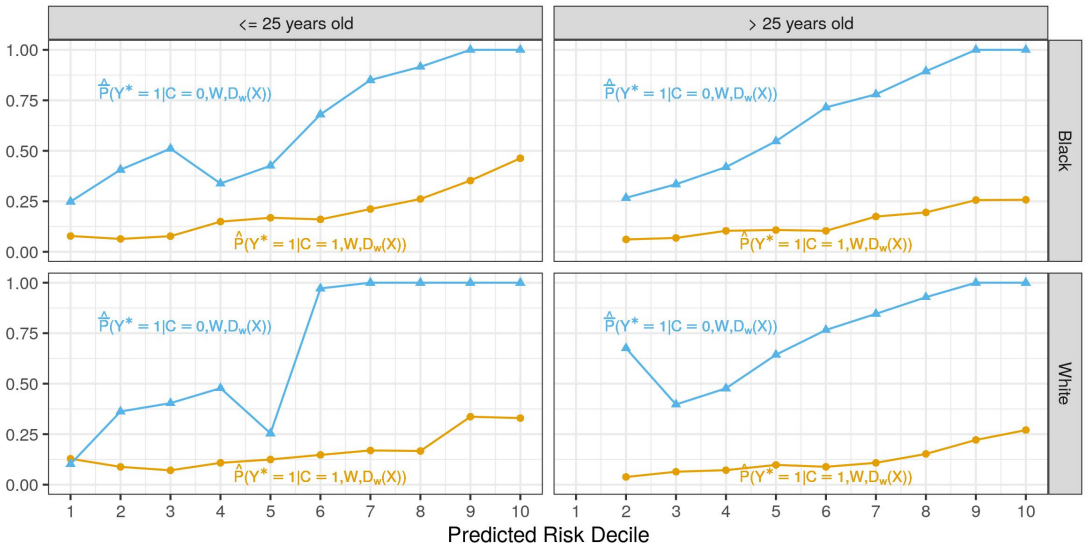
over all cases in the main estimation sample, where ϕ_t are court-by-time fixed effects.[36] I estimate

[35] There are two relevant features of the pretrial release system in New York City that suggest judges are as-if randomly assigned to cases (Leslie and Pope, 2017; Kleinberg et al., 2018*a*; Arnold, Dobbie and Hull, 2020*b*). First, bail judges are assigned to shifts in each of the five county courthouses in New York City 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.

[36] In estimating these fixed effects regressions, I follow the empirical specification in Arnold, Dobbie and Hull

the relevant quantities by taking the estimated coefficients $\hat{\beta}^{c}_{w,d,\tilde{z}}$, $\hat{\beta}^{c,y^*}_{w,d,\tilde{z}}$ and adding it to the average of the respective fixed effects associated with cases heard by the judge within each $W = w$, $D_w(X) = d$ cell.

Figure 1: Observed failure to appear rate among released defendants and constructed 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 observed 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. The judge leniency instrument $Z \in \mathcal{Z}$ is defined as the assigned judge's quintile of the constructed, leave-one-out leniency measure. Judges in New York City are quasi-randomly assigned to defendants within court-by-time cells. The bounds on the failure to appear rate among detained defendants (blue, triangles) are constructed using the most lenient quintile of judges, and by applying the instrument bounds for a quasi-random instrument (see Appendix D.1). Section 5.3.2 describes the estimation details for these bounds. Source: Rambachan and Ludwig (2021).

Figure 1 plots the observed failure to appear rate among defendants released by the judge that heard the most cases in the sample period, as well as the bounds on the failure to appear rate among detained defendants associated with the most lenient quintile of judges. The bounds are plotted 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 preferences, all released defendants have a lower observed probability of failing to appear in court (orange, circles) than the estimated upper bound on the failure to appear rate of all detained defendants (blue, triangles). Appendix Figure A1 plots the analogous quantities for each race-by-felony cell.

(2020*b*), who estimate analogous linear regressions to construct estimates of race-specific release rates and pretrial misconduct rates among released defendants.

Supplement I reports findings using an alternative empirical strategy for constructing bounds on the conditional failure to appear rate among detained defendants, which constructs bounds on the conditional failure to appear rate among detained defendants using the observed failure to appear rate among released defendants.

5.4 What Fraction of Judges Make Systematic Prediction Mistakes?

By constructing the observed failure to appear rate among released defendants and bounds on the failure to appear rate among detained defendants as in Figure 1 for each judge in the top 25, I test whether their release decisions satisfy the implied revealed preference inequalities across constructed deciles of predicted failure to appear risk (Corollary 3.4). I test the moment inequalities that compare the observed failure to appear rate among released defendants in the top half of the predicted failure to appear risk distribution against the bounds on the failure to appear rate among detained defendants in the bottom half of the predicted failure to appear risk distribution. The number of true rejections of these implied revealed preference 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 their preferences satisfy the specified exclusion restriction.

I construct the variance-covariance matrix of the observed failure to appear rates among released defendants and the bounds on the failure to appear rate among detained defendants using the empirical bootstrap conditional on the payoff-relevant characteristics W, predicted risk decile $D_w(X)$ 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 estimates of the moments and the variance-covariance matrix and has desirable power properties.

Table 1 summarizes the results from testing the implied revealed preference inequalities for each judge in the top 25 under various exclusion restrictions on their preferences. After correcting for multiple hypothesis testing, the implied revealed preference inequalities are rejected for at least 20% of judges. This provides a lower bound on the number of true rejections of the implied revealed preference inequalities among the top 25 judges.[37] For example, when both race and age are allowed to directly affect judges' preferences, violations of the implied revealed preference inequalities means that the observed release decisions could not have been generated by any possible discrimination based on the defendant's race and age nor variation in private information across defendants. Table 1 is therefore compelling evidence that a substantial fraction of judges make

[37]The number of rejections returned by a procedure that controls the family-wise error rate at the α-level provides a valid $1-\alpha$ lower bound on the number of true rejections. Given $j = 1, \ldots, m$ null hypotheses, let k be the number of false null hypotheses and let $\hat{k}$ be the number of rejections by a procedure that controls the family-wise error rate at the α-level. Observe that $P(\hat{k} \leq k) = 1 - P(\hat{k} > k)$. Since $\{\hat{k} > k\} \subseteq \{\text{at least one false rejection}\}$, $P(\hat{k} > k) \leq P(\text{at least one false rejection})$, which implies $P(\hat{k} \leq k) \geq 1 - P(\text{at least one false rejection}) \geq 1 - \alpha$.

Table 1: Estimated lower bound on the 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 $U(c, y^*; w)$			
	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 the release decisions of each judge in the top 25 are consistent with expected utility maximization at strict preference utility functions $U(c, y^*; w)$ 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. Bounds on the failure to appear rate among detained defendants are constructed using the judge leniency instrument (see Section 5.3.2). The unadjusted rejection rate reports the fraction of judges in the top 25 whose pretrial release decisions violate the moment inequalities in Corollary 3.4 at the 5% level using the conditional least-favorable hybrid test (see Section 5.4). The adjusted rejection rate reports the fraction of rejections after correcting for multiple hypothesis testing using the Holm-Bonferroni step down procedure, which controls the family-wise error rate at the 5% level. Source: Rambachan and Ludwig (2021).

prediction mistakes about failure to appear risk given defendant characteristics.

Kleinberg et al. (2018*a*) analyze whether judges in New York City make prediction mistakes in their pretrial release decisions by directly comparing the choices of all judges against those that would be made by an estimated, machine learning based decision rule (i.e., what the authors call a "reranking policy"). By comparing the choices of all judges against the model, Kleinberg et al. (2018*a*) is limited to making statements about average decision making across judges under two strong assumptions: first, that judges' preferences do not vary based on observable characteristics, and second, that preferences do not vary across judges. In contrast, I conduct my analysis judge-by-judge, allow judge preferences to flexibly vary based on defendant characteristics, and place no restrictions about how preferences vary across judges.

5.4.1 Extensions to Baseline Empirical Implementation

Robustness to the Pretrial Misconduct Outcome: I defined the outcome $Y^* \in \{0, 1\}$ to be whether a defendant would fail to appear in court if they were released. If this is incorrectly specified and a judge's preferences depend on some other definition of pretrial misconduct, then rejections of expected utility maximization may be driven by mis-specification of the outcome. For example, even though the New York City pretrial release system explicitly asks judges to only consider failure to appear risk, judges may additionally base their release decisions on whether a defendant would be arrested for a new crime (Arnold, Dobbie and Yang, 2018; Kleinberg et al., 2018*a*; Arnold, Dobbie and Hull, 2020*b*). I reconduct my empirical analysis defining the outcome to be whether the defendant would either fail to appear in court or be re-arrested for a new

crime ("any pre-trial misconduct"). Appendix Table A1 shows that, for this alternative definition of pretrial misconduct, the pretrial release decisions of at least 64% of judges in New York City are inconsistent with expected utility maximization at preferences that satisfy these conjectured exclusion restrictions and accurate beliefs about any pretrial misconduct risk given defendant characteristics.

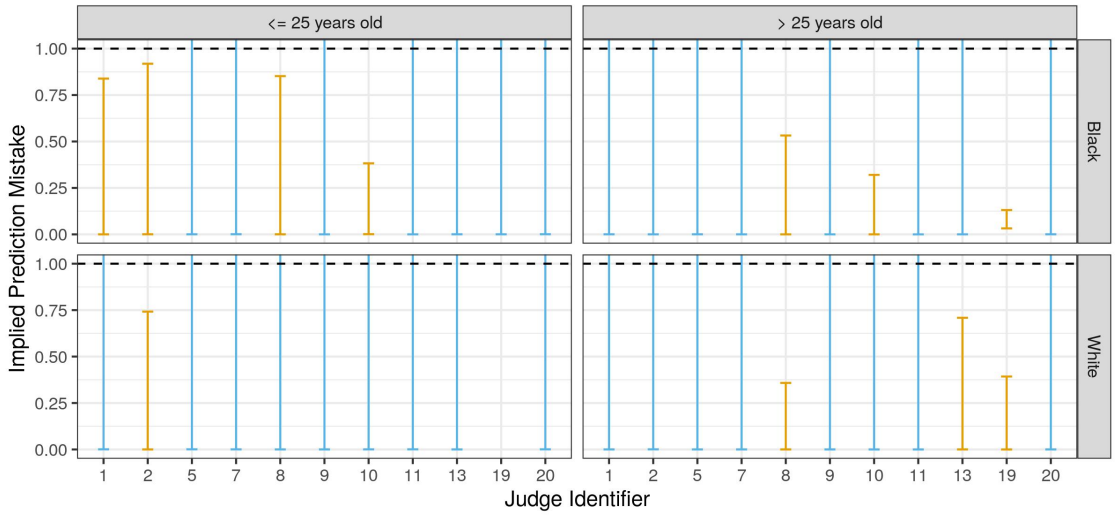
Robustness to Pretrial Release Definition: My empirical analysis collapsed the pretrial release decision into a binary choice of either to release or detain a defendant. In practice, judges in New York City decide whether to release a defendant "on recognizance," meaning the defendant is released automatically without bail conditions, or set monetary bail conditions. Consequently, judges could be making two distinct prediction mistakes: first, judges may be systematically mis-predicting failure to appear risk; and second, judges may be systematically mis-predicting the ability of defendants to post a specified bail amount.

In Supplement I.4, I redefine the judge's choice to be whether or not to release the defendant on recognizance and the outcome as both whether the defendant will pay the specified bail amount and whether the defendant would fail to appear in court if released. Since the modified outcome takes on multiple values, I use Theorem 2.1 to show that expected utility maximization at accurate beliefs about bail payment ability and failure to appear risk can again characterized as a system of moment inequalities (Proposition I.1). I find that at least 32% of judges in New York City 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 given defendant characteristics.

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 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 violate the implied revealed preference inequalities, I construct a 95% confidence interval for their implied prediction mistakes $\delta(w,d)/\delta(w,d')$ 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 $\tau(w,d), \tau(w,d')$ at the bottom and top deciles of the predicted failure to appear risk distribution using test inversion based on Theorem 4.2. I then construct a 95% confidence interval for the implied prediction mistake by calculating $\frac{(1-\tau(w,d))/\tau(w,d)}{(1-\tau(w,d'))/\tau(w,d')}$ for each pair $\tau(w,d), \tau(w,d')$ of values that lie in the joint confidence set as in Corollary 4.1.

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 $\delta(w,d)/\delta(w,d')$ 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. The implied prediction mistake $\delta(w,d)/\delta(w,d')$ measures the degree to which judges' beliefs underreact or overreact to variation in failure to appear risk. 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). These confidence intervals are constructed by first constructing a 95% joint confidence interval for a judge's reweighed utility threshold $\tau(w,d), \tau(w,d')$ using test inversion based on the moment inequalities in Theorem 4.2, and then constructing the implied prediction mistake $\delta(w,d)/\delta(w,d')$ associated with each pair $\tau(w,d), \tau(w,d')$ in the joint confidence set (Corollary 4.1). See Section 4.2 for theoretical details on the implied prediction mistake and Section 5.5 for the estimation details. Source: Rambachan and Ludwig (2021).

Figure 2 plots the constructed confidence intervals for the implied prediction mistakes $\delta(w,d)/\delta(w,d')$ for each judge over the race-and-age cells. Whenever informative, the confidence intervals highlighted in orange lie everywhere below one, indicating that these judges' are acting as-if their implied beliefs about failure to appear risk underreact to predictable variation in failure to appear risk. That is, 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 "over-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 (Gabaix, 2014; Caplin and Dean, 2015; Gabaix, 2019). For example, perhaps judges are failing to pay attention to characteristics that distinguish predictably low and predictable high failure to appear risk defendants, resulting in their implied predictions to underreact to variation in failure

to appear risk. Developing formal tests for these behavioral mechanisms in empirical settings like pretrial release is beyond the scope of this paper. Analogous estimates for race-and-felony charge cells are summarized in Figure A2.

Supplement I 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.

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 also report the cells of defendants on which the maximum studentized violation of the revealed preference inequalities in Corollary 3.4 occurs. This shows which defendants are associated with the largest violations of the revealed preference inequalities.

Table 2: Location of the maximum studentized violation of revealed preference inequalities among judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs about failure to appear risk given defendant characteristics.

	Utility Functions $U(c, y; w)$	
	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 maximum studentized violation of revealed preference inequalities in Corollary 3.4 among judges whose release decisions are inconsistent with expected utility maximization at accurate beliefs and preferences that depend on both the defendant's race and age as well as the defendant's race and whether the defendant was charged with a felony. Bounds on the failure to appear rate among detained defendants are constructed using the judge leniency instrument (see Section 5.3.2). Among judge's whose release decision violate the revealed preference inequalities at the 5% level, I report the fraction of judges for whom the maximal studentized violation 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. Source: Rambachan and Ludwig (2021).

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 violation 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. Consistent with the previous findings

that judges' implied beliefs underreact to variation in failure to appear risk, I find all maximal violations of the revealed preference inequalities occur over defendants that lie in the tails of the predicted risk distribution, and furthermore the majority occur over black defendants at the tails of the predicted risk distribution. Taken together, these empirical findings highlight that detectable prediction mistakes primarily occur on defendants at the tails of the predicted risk distribution.

6 The Effects of Algorithmic Decision-Making

Finally, I illustrate that the preceding behavioral analysis has important policy implications about the design of algorithmic decision systems by analyzing various policy counterfactuals that replace judges with algorithmic decisions rules in the New York City pretrial release setting. As a technical step, I show that average social welfare under a candidate decision rule is partially identified due to the missing data problem, and provide simple methods for inference on this quantity.

6.1 Social Welfare Under Candidate Decision Rules

Focusing on binary screening decisions, consider a policymaker whose preferences over the unknown outcome and choices are summarized by the *social welfare function* $U^*(0,0) < 0, U^*(1,1) < 0$. The policymaker evaluates a candidate decision rule $p^*(w,x) \in [0,1]$, which denotes the probability $C = 1$ is chosen given $W = w, X = x$. Due to the missing data problem, expected social welfare under the candidate decision rule is partially identified. I characterize the identified set of expected social welfare under the candidate decision rule, and show that testing the null hypothesis that expected social welfare is equal to some value is equivalent to testing a series of moment inequalities with nuisance parameters that enter linearly. These results extend to analyzing expected social welfare under the decision maker's observed choices (see Appendix D.3).

For a candidate decision rule $p^*(w,x)$, expected social welfare at $(w,x) \in \mathcal{W} \times \mathcal{X}$ is

$$\theta(w,x) = \ell(w,x;p^*,U^*)P_{Y^*}(1 \mid w,x) + \beta(w,x;p^*,U^*), \tag{6}$$

where $\ell(w,x;p^*,U^*) := U^*(1,1)p^*(w,x) - U^*(0,0)(1-p^*(w,x))$ and $\beta(w,x;p^*,U^*) := U^*(0,0)(1-p^*(w,x))$. Total expected social welfare equals

$$\theta(p^*,U^*) = \beta(p^*,U^*) + \sum_{(w,x)\in\mathcal{W}\times\mathcal{X}} P(w,x)\ell(w,x;p^*,U^*)P_{Y^*}(1 \mid w,x) \tag{7}$$

where $\beta(p^*,U^*) := \sum_{(w,x)\in\mathcal{W}\times\mathcal{X}} P(w,x)\beta(w,x;p^*,U^*)$. This definition of the social welfare function is strictly utilitarian. It does not incorporate additional fairness considerations that have received much attention in an influential literature in computer science are particularly important in the criminal justice system (e.g., see Barocas and Selbst, 2016; Mitchell et al., 2019; Barocas,

Hardt and Narayanan, 2019; Chouldechova and Roth, 2020).[38]

Since $P_{Y^*}(1 \mid w, x)$ is partially identified due to the missing data problem, total expected social welfare is also partially identified and its sharp identified set of total expected welfare is an interval.

Proposition 6.1. *Consider a binary screening decision, a policymaker with social welfare function $U^*(0,0) < 0, U^*(1,1) < 0$ and a candidate decision rule $p^*(w,x)$. The sharp identified set of total expected social welfare, denoted by $\mathcal{H}_P(\theta(p^*, U^*); \mathcal{B})$, is an interval with $\mathcal{H}_P(\theta(p^*, U^*); \mathcal{B}) = \left[\underline{\theta}(p^*, U^*), \overline{\theta}(p^*, U^*)\right]$, where*

$$\underline{\theta}(p^*, U^*) = \beta(p^*, U^*) + \left\{ \begin{array}{ll} \min\limits_{\left\{\tilde{P}_{Y^*}(\cdot|w,x):\ (w,x)\in\mathcal{W}\times\mathcal{X}\right\}} & \sum\limits_{(w,x)\in\mathcal{W}\times\mathcal{X}} P(w,x)\ell(w,x;p^*,U^*)\widetilde{P}_{Y^*}(1 \mid w,x) \\ \textit{s.t. } \tilde{P}_{Y^*}(\cdot \mid w,x) \in \mathcal{H}_P(P_{Y^*}(\cdot \mid w,x); \mathcal{B}_{0,w,x})\ \forall (w,x) \in \mathcal{W}\times\mathcal{X} & \end{array} \right\},$$

$$\overline{\theta}(p^*, U^*) = \beta(p^*, U^*) + \left\{ \begin{array}{ll} \max\limits_{\left\{\tilde{P}_{Y^*}(\cdot|w,x):\ (w,x)\in\mathcal{W}\times\mathcal{X}\right\}} & \sum\limits_{(w,x)\in\mathcal{W}\times\mathcal{X}} P(w,x)\ell(w,x;p^*,U^*)\widetilde{P}_{Y^*}(1 \mid w,x) \\ \textit{s.t. } \tilde{P}_{Y^*}(\cdot \mid w,x) \in \mathcal{H}_P(P_{Y^*}(\cdot \mid w,x); \mathcal{B}_{0,w,x})\ \forall (w,x) \in \mathcal{W}\times\mathcal{X} & \end{array} \right\}.$$

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 candidate decision rule and joint distribution of the characteristics (W, 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 6.2. *Consider a binary screening decision, a policymaker with social welfare function $U^*(0,0) < 0, U^*(1,1) < 0$ and a known candidate decision rule $p^*(w,x)$. Conditional on the characteristics (W, X), testing the null hypothesis $H_0\colon \theta(p^*, U^*) = \theta_0$ is equivalent to testing whether*

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

where $\ell(p^*, U^*)$ is the $d_w d_x$-dimensional vector with elements $P(w,x)\ell(w,x;p^*,U^*)$, $P^{c=1,y^*=1}$ is the $d_w d_x$-dimensional vector of moments $P_{C,Y^*}(1,1 \mid w,x)$, $\underline{P}^{c=0,y^*=1}, \overline{P}^{c=0,y^*=1}$ are the $d_w d_x$-dimensional vectors of lower and upper bounds on $P_{C,Y^*}(0,1 \mid w,x)$ respectively, and $\widetilde{A}$ is a known matrix.[39]

[38] These results could be extended to incorporate a social welfare function varies across groups defined by the characteristics as in Rambachan et al. (2021), or a penalty that depends on the composition of the individuals that receive $C = 1$ as in Kleinberg et al. (2018*b*).

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

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. Using this testing reduction requires that the candidate decision rule be known, which can be achieved by constructing the decision rule on held-out data.

6.2 Measuring the Effects of Algorithms in Pretrial Release Decisions

I use these results to compare total expected social welfare under the observed decisions of judges in New York City against total expected social welfare under counterfactual algorithmic decision rules. I vary the relative cost of detaining an individual that would not fail to appear in court $U^*(0,0)$ (i.e., an "unnecessary detention"), normalizing $U^*(1,1) = -1$. For a particular choice of the social welfare function, I construct an algorithmic decision rule that decides whether to release individuals based on a prediction of the probability of pretrial misconduct at each possible cell of payoff relevant characteristics W and each decile of predicted failure to appear risk $D_w(X)$. The decision rule is a threshold rule, whose cutoff depends on the particular parametrization of the social welfare function. Appendix C.4 discusses the construction of this decision rule.

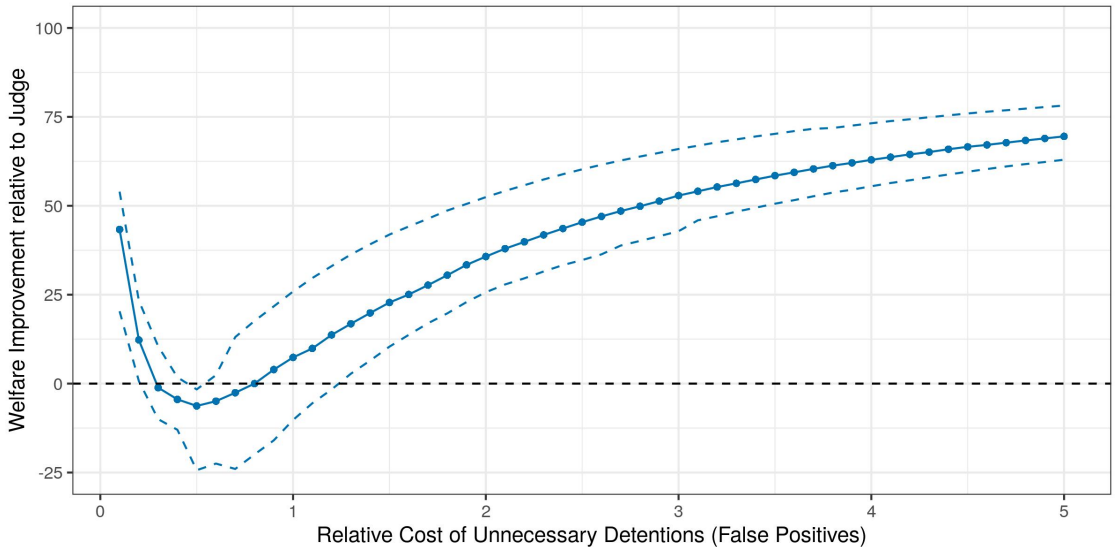
I construct 95% confidence intervals for total expected social welfare under the algorithmic decision rule and the judge's observed released decisions. I report the ratio of worst-case total expected social welfare under the algorithmic decision rule against worst-case total expected social welfare under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

6.2.1 Automating Judges Who Make Detectable Prediction Mistakes

I compare the algorithmic decision rule against the release decisions of judges who were found to make detectable prediction mistakes. Figure 3 plots the improvement in worst-case total expected social welfare under the algorithmic decision rule that fully replaces these judges over all decisions against the observed release decisions of these judges. For most values of the social welfare function, worst-case total expected social welfare under the algorithmic decision rule is strictly larger than worst-case total expected social welfare under these judges' decisions. Recall these judges primarily made detectable prediction mistakes over defendants in the tails of the predicted failure to appear risk distribution. Over the remaining defendants, however, their choices were consistent with expected utility maximization at accurate beliefs about failure to appear risk. Consequently,

the difference in expected social welfare under the algorithmic decision rule and the judges' decisions are driven by three forces: first, the algorithmic decision rule corrects detectable prediction mistakes over the tails of the predicted risk distribution; second, the algorithmic decision rule corrects possible misalignment between the policymaker's and judges' preferences over the remaining defendants; and third, the judges may observe predictive private information over the remaining defendants that is unavailable to the algorithmic decision rule.

Figure 3: Ratio of total expected social welfare under algorithmic decision rule relative to release decisions of judges that make detectable prediction mistakes about failure to appear risk.

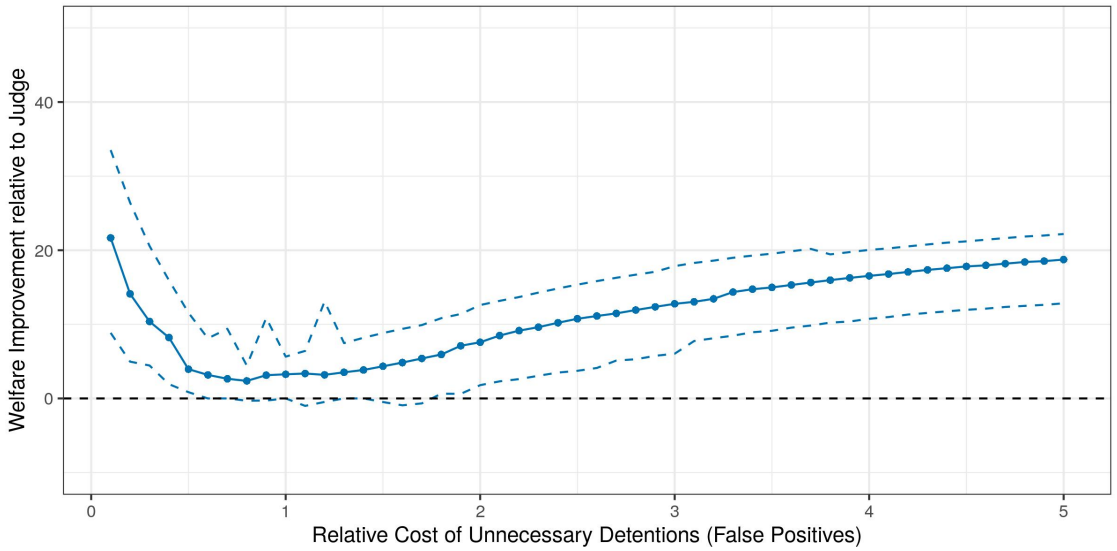


Notes: This figure reports the change in worst-case total expected social welfare under the algorithmic decision rule that fully automates decision-making against the observed release decisions of judges who were found to make detectable prediction mistakes. Worst case total expected social welfare under each decision rule is computed by constructing 95% confidence intervals for total expected social welfare under the decision rule, and reporting smallest value that lies in the confidence interval. These decisions rules are constructed and evaluated over race-by-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 $U^*(0,0)$ (i.e., an unnecessary detention). The solid line plots the median change across judges, and the dashed lines report the minimum and maximum change across judges. See Section 6.2 for further details. Source: Rambachan and Ludwig (2021).

For social welfare costs of unnecessary detentions ranging over $U^*(0,0) \in [0.3, 0.8]$, the algorithmic 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 total expected social welfare by the race of the defendant, highlighting that these costs are particularly large over white defendants. At these values, judges' preferences may be sufficiently aligned with the policymaker and observe sufficiently predictive private information over the remaining defendants that it is costly to fully automate their decisions. 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 judges' decisions dominate the algorithmic decision rule.

Figure 4: Ratio of total expected social welfare under algorithmic decision rule that corrects prediction mistakes relative to release decisions of judges that make detectable prediction mistakes about failure to appear risk.



Notes: This figure reports the change in worst-case total expected social welfare under the algorithmic decision rule that only replaces the judge on cases in the tails of the predicted failure to appear risk distribution (deciles 1-2, and deciles 9-10) against the judge's observed release decisions. Worst case total expected social welfare under each decision rule is computed by constructing 95% confidence intervals for total expected social welfare under the decision rule, and reporting smallest value that lies in the confidence interval. These decisions rules are constructed and evaluated over race-by-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 $U^*(0,0)$ (i.e., an unnecessary detention). The solid line plots the median change across judges, and the dashed lines report the minimum and maximum change across judges. See Section 6.2 for further details. Source: Rambachan and Ludwig (2021).

The 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 detectable prediction mistakes. I 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 is a common way of statistical risk assessments are used in pretrial release systems throughout the United States (Stevenson, 2018; Albright, 2019; Dobbie and Yang, 2019). This algorithmic decision rule only corrects the detectable prediction mistakes, and its welfare effects are plotted in Figure 4. The algorithmic decision rule that only corrects 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 provides a behavioral mechanism for recent machine learning methods that attempt to estimate whether a decision should be made by an algorithm or instead deferred to a decision maker (e.g., Madras, Pitassi and Zemel, 2018; Raghu et al., 2019; Wilder, Horvitz and Kamar, 2020). These results show that removing judicial discretion over cases where detectable prediction mistakes are made but otherwise deferring to them over all other cases may be a "free lunch." Deciding whether to automate a decision or defer to a decision maker therefore requires understanding whether the decision maker makes systematic prediction mistakes, and if so on what decisions.

6.2.2 Automating Judges Who Do Not Make Detectable 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. Automating these judge's release decisions may strictly lower worst-case expected social welfare for a range of social welfare costs of unnecessary detentions, These judges are making pretrial release decisions as-if their preferences 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 plots the results by defendant race and Figure A7 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 assessing the value of their private information against the degree to which they are misaligned, which is beyond the scope of this paper.[40]

7 Conclusion

This paper develops an econometric framework for testing whether a decision maker makes prediction mistakes in high stakes, empirical settings such as hiring, medical testing and treatment, and pretrial release. I characterized expected utility maximization, 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 pretrial release system in New York City, 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 counterfactual social welfare under alternative algorithmic release rules against the observed release

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

decisions of judges.

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

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

Arnold, David, Will Dobbie, and Peter Hull. 2020*b*. "Measuring Racial Discrimination in Bail Decisions." NBER Working Paper No. 26999.

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

Augenblick, Ned, and Eben Lazarus. 2020. "Restrictions on Asset-Price Movements Under Rational Expectations: Theory and Evidence."

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.

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.

Bergemann, Dirk, Benjamin Brooks, and Stephen Morris. 2019. "Counterfactuals with Latent Information."

Bergemannn, Dirk, and Stephen Morris. 2013. "Robust Predictions in Games with Incomplete Information." *Econometrica*, 81(4): 1251–1308.

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

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

Berk, Richard A., Susan B. Sorenson, and Geoffrey Barnes. 2016. "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions." *Journal of Empirical Legal Studies*, 13(1): 94–115.

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, Yueran Ma, and Andrei Shleifer. 2020. "Overreaction in Macroeconomic Expectations." *American Economic Review*, 110(9): 2748–82.

Bugni, Federico A., Ivan A. Canay, and XiaoXia Shi. 2015. "Specification tests for partially identified models defined by moment inequalities." *Journal of Econometrics*, 185: 259–282.

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.

Campbell, John Y. 2003. "Chapter 13 Consumption-based asset pricing." In *Financial Markets and Asset Pricing*. Vol. 1 of *Handbook of the Economics of Finance*, 803–887. Elsevier.

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. 2016. "Measuring and Modeling Attention." *Annual Review of Economics*, 8: 379–403.

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

Caplin, Andrew, and Daniel Martin. 2021. "Comparison of Decisions under Unknown Experiments."

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

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

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.

Chambers, Christopher P., Ce Liu, and Seung-Keun Martinez. 2016. "A Test for Risk-Averse Expected Utility." *Journal of Economic Theory*, 163: 775–785.

Chan, David C., Matthew Gentzkow, and Chuan Yu. 2020. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." NBER Working Paper No. 26467.

Chandra, Amitabh, and Douglas O. Staiger. 2007. "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks." *Journal of Political Economy*, 115(1): 103–140.

Chandra, Amitabh, and Douglas O. Staiger. 2020. "Identifying Sources of Inefficiency in Healthcare." *The Quarterly Journal of Economics*, 135(2): 785—-843.

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.

Chernozhukov, Victor, Sokbae Lee, and Adam M. Rosen. 2013. "Intersection Bounds: Estimation and Inference." *Econometrica*, 81(2): 667–737.

Cho, JoonHwan, and Thomas M. Russell. 2020. "Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments."

Chouldechova, Alexandra, and Aaron Roth. 2020. "A Snapshot of the Frontiers of Fairness in Machine Learning." *Communications of the ACM*, 63(5): 82–89.

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.

Cochrane, John H. 2011. "Presidential Address: Discount Rates." *The Journal of Finance*, 66(4): 1047–1108.

Coston, Amanda, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. "Counterfactual Risk Assessments, Evaluation and Fairness." *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 582—593.

Coston, Amanda, Ashesh Rambachan, and Alexandra Chouldechova. 2021. "Characterizing Fairness Over the Set of Good Models Under Selective Labels."

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.

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

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-Arteaga, Maria, Artur Dubrawski, and Alexandra Chouldechova. 2021. "Leveraging Expert Consistency to Improve Algorithmic Decision Support." arXix preprint, arXiv:2101.09648.

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

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

D'Haultfoeuille, Xavier, Christophe Gaillac, and Arnaud Maurel. 2020. "Rationalizing Rational Expectations: Characterization and Tests." arXiv preprint, arXiv:2003.11537.

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. 2020. "Measuring Bias in Consumer Lending."

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. 2020. "New Developments in Revealed Preference Theory: Decisions Under Risk, Uncertainty, and Intertemporal Choice." *Annual Review of Economics*, 12: 299–316.

Echenique, Federico, and Kota Saito. 2015. "Savage in the Market." *Econometrica*, 83(4): 1467–1495.

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.

Elliott, Graham, Allan Timmerman, and Ivana Komunjer. 2005. "Estimation and Testing of Forecast Rationality under Flexible Loss." *The Review of Economic Studies*, 72(4): 1107–1125.

Elliott, Graham, Ivana Komunjer, and Allan Timmerman. 2008. "Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss?" *Journal of the European Economic Association*, 6(1): 122–157.

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

Fang, Zheng, Andres Santos, Azeem M. Shaikh, and Alexander Torgovitsky. 2020. "Inference for Large-Scale Linear Systems with Known Coefficients."

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. "Certifying and Removing Disparate Impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

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. 2018. "Predictably Unequal? The Effects of Machine Learning on Credit Markets."

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.

Gafarov, Bulat. 2019. "Inference in high-dimensional set-identified affine models." arXiv preprint, arXiv:1904.00111.

Gelbach, Jonah. 2021. "Testing Economic Models of Discrimination in Criminal Justice."

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

Gennaioli, Nicola, Yueran Ma, and Andrei Shleifer. 2016. "Expectations and Investment." *NBER Macroeconomics Annual*, 30: 379–431.

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

Green, Ben, and Yiling Chen. 2019*a*. "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments." *FAT* '19*, 90–99. New York, NY, USA:Association for Computing Machinery.

Green, Ben, and Yiling Chen. 2019*b*. "The Principles and Limits of Algorithm-in-the-Loop Decision Making." *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

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.

Gul, Faruk, and Wolfgang Pesendorfer. 2006. "Random Expected Utility." *Econometrica*, 74(1): 121–146.

Gul, Faruk, Paulo Natenzon, and Wolfgang Pesendorfer. 2014. "Random Choice as Behavioral Optimization." *Econometrica*, 82: 1873–1912.

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.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3323–3331.

Heckman, James J. 1974. "Shadow Prices, Market Wages, and Labor Supply." *Econometrica*, 42(4): 679–694.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*, 47(1): 153–161.

Heckman, James J., and Edward J. Vytlacil. 2006. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics*. Vol. 6, 4779–4874.

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

Hilgard, Sophie, Nir Rosenfeld, Mahzarin R Banaji, Jack Cao, and David Parkes. 2021. "Learning Representations by Humans, for Humans." Vol. 139 of *Proceedings of Machine Learning Research*, 4227–4238.

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

Ho, Kate, and Adam M. Rosen. 2017. "Partial Identification in Applied Research: Benefits and Challenges." *Advances in Economics and Econometrics: Eleventh World Congress*, , ed. Bo Honoré, Ariel Pakes, Monika Piazzesi and Larry Samuelson Vol. 2, 307–359. Cambridge University Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81: 945–960.

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.

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62: 467–475.

Jacob, Brian A., and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics*, 26(1): 101–136.

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

Jung, Jongbin, Ravi Shroff, Avi Feller, and Sharad Goel. 2020*b*. "Bayesian Sensitivity Analysis for Offline Policy Evaluation." *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 64–70.

Kallus, Nathan, and Angela Zhou. 2018. "Confounding-Robust Policy Improvement." *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.

Kallus, Nathan, Xiaojie Mao, and Angela Zhou. 2018. "Interval Estimation of Individual-Level Causal Effects Under Unobserved Confounding." arXiv preprint arXiv:1810.02894.

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.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. "Consumer credit-risk models via machine-learning algorithms." *Journal of Banking & Finance*, 34(11): 2767 – 2787.

Kitagawa, Toru. 2020. "The Identification Region of the Potential Outcome Distributions under Instrument Independence." Cemmap Working Paper CWP23/20.

Kitamura, Yuichi, and Jorg Stoye. 2018. "Nonparametric Analysis of Random Utility Models." *Econometrica*, 86(6): 1883–1909.

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.

Kubler, Felix, Larry Selden, and Xiao Wei. 2014. "Asset Demand Based Tests of Expected Utility Maximization." *American Economic Review*, 104(11): 3459–3480.

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.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. "Delayed Impact of Fair Machine Learning." *Proceedings of the 35th International Conference on Machine Learning*.

Lu, Jay. 2016. "Random Choice and Private Information." *Econometrica*, 84(6): 1983–2027.

Lu, Jay. 2019. "Bayesian Identification: A Theory for State-Dependent Utilities." *American Economic Review*, 109(9): 3192–3228.

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

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.

Manski, Charles F. 2004. "Measuring Expectations." *Econometrica*, 72(5): 1329–1376.

Manski, Charles F. 2017. "Improving Clinical Guidelines and Decisions Under Uncertainty." NBER Working Paper No. 23915.

Marquardt, Kelli. 2021. "Mis(sed) Diagnosis: Physician Decision Making and ADHD."

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.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2019. "Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions." arXiv Working Paper, arXiv:1811.07867.

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. 2020. "Who is tested for heart attack and who should be: predicting patient risk and physician error." NBER Working Paper Series, Working Paper No. 26168.

Natenzon, Paulo. 2019. "Random Choice and Learning." *Journal of Political Economy*, 127(1): 419–457.

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.

Polisson, Matthew, John K. H. Quah, and Ludovic Renou. 2020. "Revealed Preferences over Risk and Uncertainty." *American Economic Review*, 110(6): 1782–1820.

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. URL: https://urbanlabs.uchicago.edu/projects/empirical-analysis-of-prediction-mistakes.

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.

Ribers, Michael Allan, and Hannes Ullrich. 2020. "Machine Predictions and Human Decisions with Variation in Payoffs and Skills."

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2011. "Can You Recognize an Effective Teacher When You Recruit One?" *Education Finance and Policy*, 6(1): 43–74.

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

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66(5): 688–701.

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

Russell, Thomas M. 2019. "Sharp Bounds on Functionals of the Joint Distribution in the Analysis of Treatment Effects." *Journal of Business & Economic Statistics*.

Simon, Herbert A. 1955. "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics*, 69(1): 99–118.

Simon, Herbert A. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review*, 63(2): 129–138.

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.

Tan, Sarah, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. "Investigating Human + Machine Complementarity for Recidivism Predictions." arXiv preprint, arXiv:1808.09123.

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

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

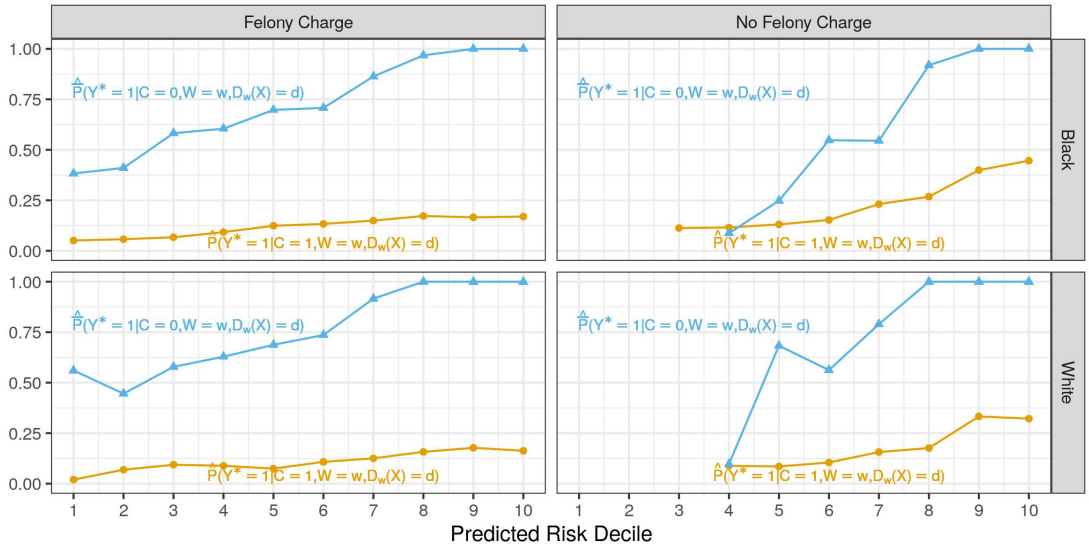
Wright, Marvin N., and Andreas Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software, Articles*, 77(1): 1–17.

Yadlowsky, Steve, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. 2020. "Bounds on the conditional and average treatment effect with unobserved confounding factors." arXiv preprint arXiv:1808.09521.

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

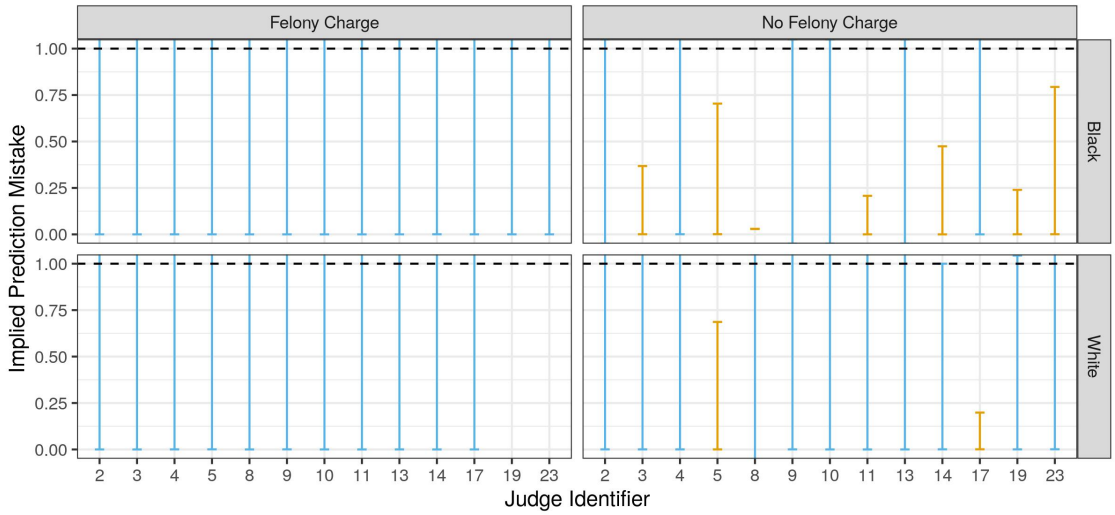
A Additional Figures and Tables

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



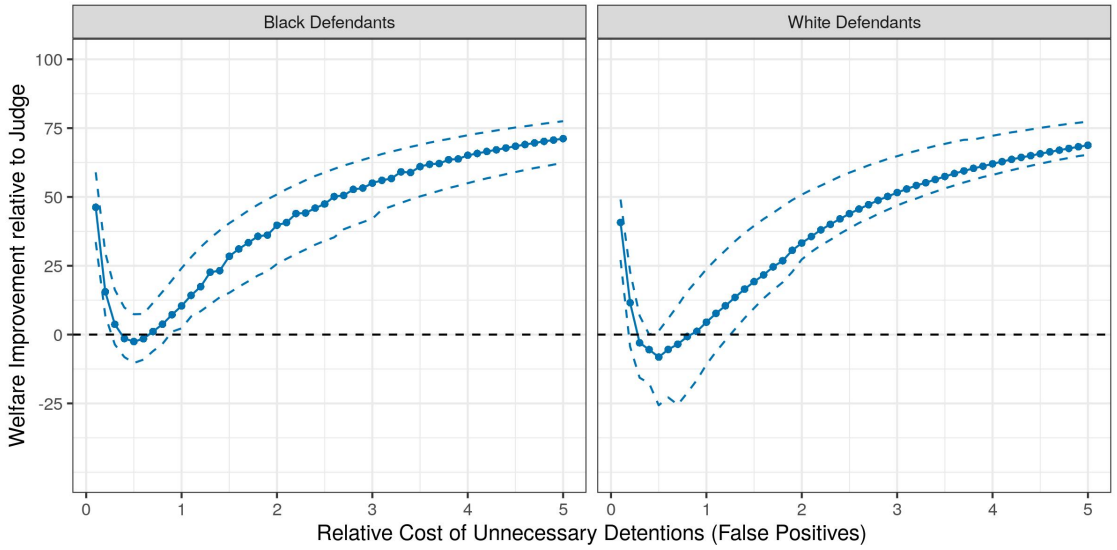
Notes: This figure plots the observed failure to appear rate among released defendants (orange, circles) and the bounds based on the judge leniency for the failure to appear rate among detained defendants (blue, triangles) at each decile of predicted failure to appear risk and race-by-felony charge cell for the judge that heard the most cases in the main estimation sample. he judge leniency instrument $Z \in \mathcal{Z}$ is defined as the assigned judge's quintile of the constructed, leave-one-out leniency measure. Judges in New York City are quasi-randomly assigned to defendants within court-by-time cells. The bounds on the failure to appear rate among detained defendants (blue, triangles) are constructed using the most lenient quintile of judges, and by applying the instrument bounds for a quasi-random instrument (see Appendix D.1). Section 5.3.2 describes the estimation details for these bounds. Source: Rambachan and Ludwig (2021).

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.



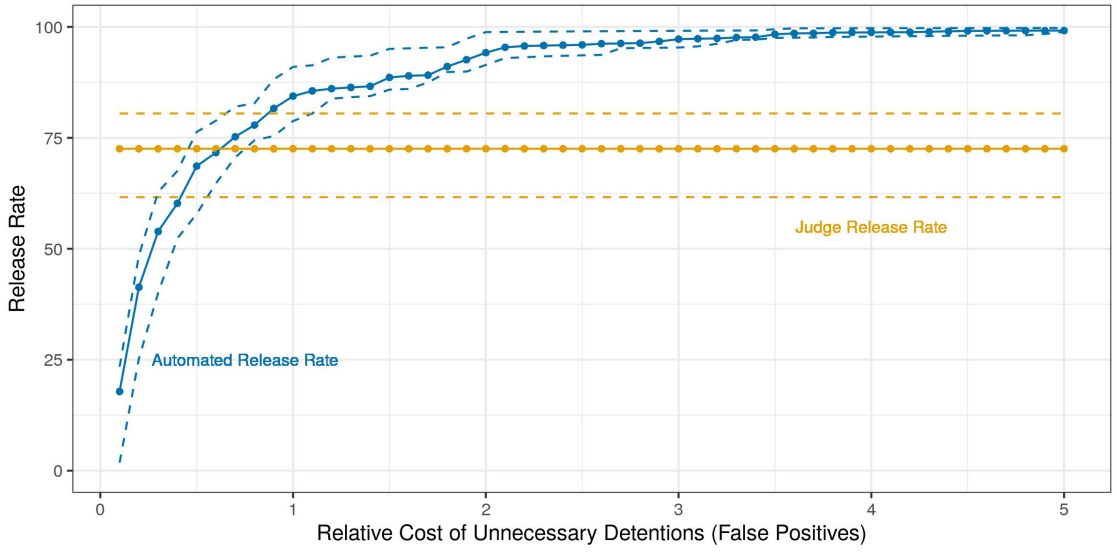
Notes: This figure plots the 95% confidence interval on the implied prediction mistake $\delta(w,d)/\delta(w,d')$ between the top decile d and bottom decile d' of the predicted failure to appear risk distribution for each judge in the top 25 whose pretrial release decisions violated the implied revealed preference inequalities (Table 1) and each race-by-felony charge cell. The implied prediction mistake $\delta(w,d)/\delta(w,d')$ measures the degree to which judges' beliefs underreact or overreact to variation in failure to appear risk. 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). These confidence intervals are constructed by first constructing a 95% joint confidence interval for a judge's reweighed utility threshold $\tau(w,d),\tau(w,d')$ using test inversion based on the moment inequalities in Theorem 4.2, and then constructing the implied prediction mistake $\delta(w,d)/\delta(w,d')$ associated with each pair $\tau(w,d),\tau(w,d')$ in the joint confidence set (Corollary 4.1). See Section 4.2 for theoretical details on the implied prediction mistake and Section 5.5 for the estimation details. Source: Rambachan and Ludwig (2021).

Figure A3: Ratio of total expected social welfare under algorithmic decision rule relative to observed decisions of judges that make detectable prediction mistakes about failure to appear risk by defendant race.



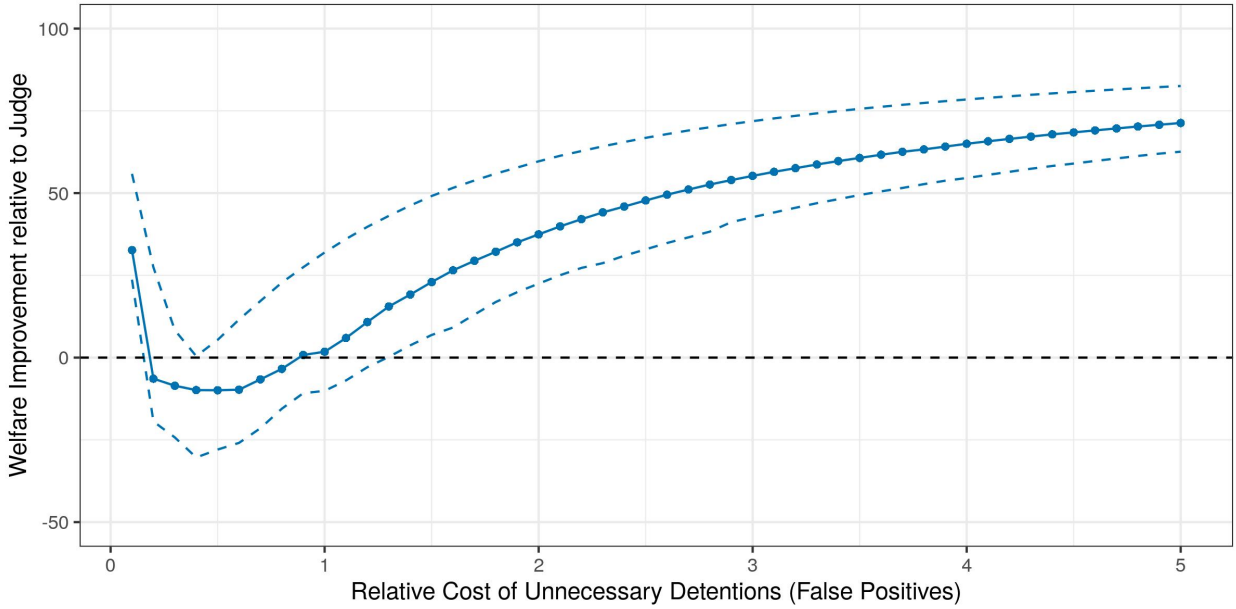
Notes: This figure reports the change in worst-case total expected social welfare under the algorithmic decision rule that fully automates decisions against the judge's observed release decisions among judges who were found to make detectable prediction mistakes, broken out by defendant race. Worst case total expected social welfare under each decision rule is computed by first constructing a 95% confidence interval for total expected social welfare under the decision rule, and reporting smallest value that lies in the confidence interval. These decisions rules are constructed and evaluated over race-by-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 $U^*(0,0)$ (i.e., an unnecessary detention). 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 6.2 for further details. Source: Rambachan and Ludwig (2021).

Figure A4: Overall release rates under algorithmic decision rule relative to the observed release rates of judges that make detectable prediction mistakes about failure to appear risk.



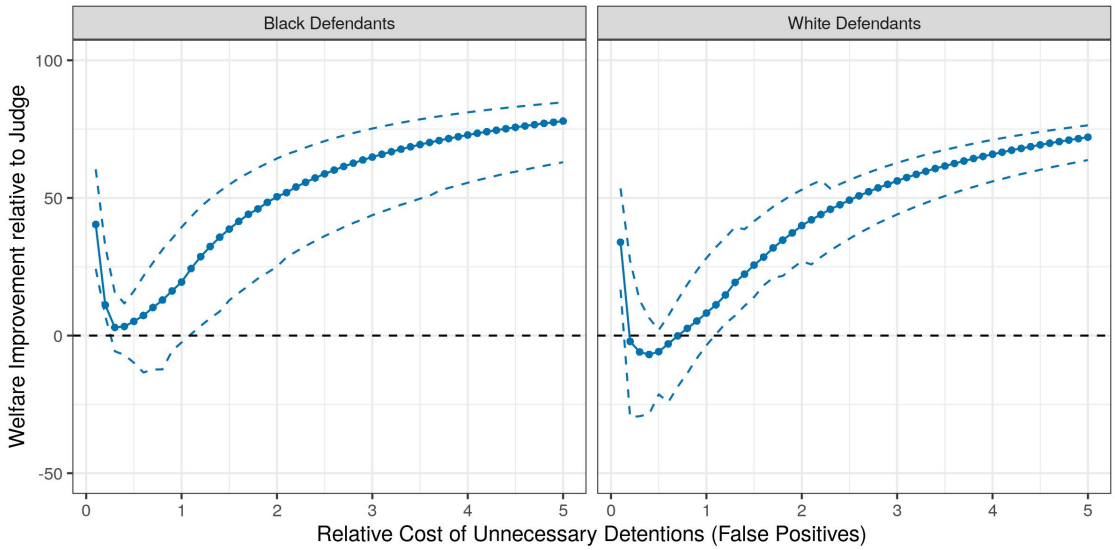
Notes: This figure reports the overall release rate of the algorithmic decision rule that fully automates decisions against the judge's observed release rates among judges who were found to make detectable prediction mistakes. These decisions rules are constructed and evaluated over race-by-age cells 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 $U^*(0,0)$ (i.e., an unnecessary detention). The solid line plots the median release rate across judges that make detectable prediction mistakes, and the dashed lines report the minimum and maximum release rates across judges. See Section 6.2 for further details. Source: Rambachan and Ludwig (2021).

Figure A5: Ratio of total expected social welfare under algorithmic decision rule relative to release decisions of judges that do not make detectable prediction mistakes about failure to appear risk.



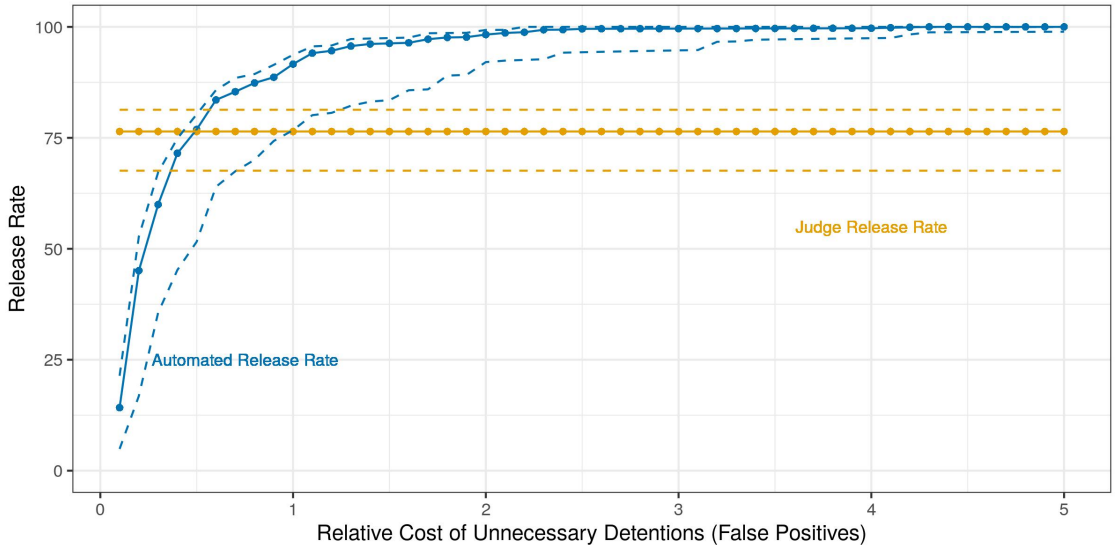
Notes: This figure reports the change in worst-case total expected social welfare under the algorithmic decision rule that fully automates decision-making against the judge's observed release decisions among judges whose choices were consistent with expected utility maximization at accurate beliefs about failure to appear risk. Worst case total expected social welfare under each decision rule is computed by constructing 95% confidence intervals for total expected social welfare under the decision rule, and reporting smallest value that lies in the confidence interval. These decisions rules are constructed and evaluated over race-by-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 $U^*(0,0)$ (i.e., an unnecessary detention). The solid line plots the median change across judges, and the dashed lines report the minimum and maximum change across judges. See Section 6.2 for further details. Source: Rambachan and Ludwig (2021).

Figure A6: Ratio of total expected social welfare under algorithmic decision rule relative to observed decisions of judges that do not make detectable prediction mistakes by defendant race.



Notes: This figure reports the change in worst-case total expected social welfare under the algorithmic decision rule that fully automates decision-making against the judge's observed release decisions among judges whose choices were consistent with expected utility maximization behavior at accurate beliefs, broken out by defendant race. Worst case total expected social welfare under each decision rule is computed by first constructing a 95% confidence interval for total expected social welfare under the decision rule, and reporting smallest value that lies in the confidence interval. These decisions rules are constructed and evaluated over race-by-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 $U^*(0,0)$ (i.e., an unnecessary detention). The solid line plots the median change across judges, and the dashed lines report the minimum and maximum change across judges. See Section 6.2 for further details. Source: Rambachan and Ludwig (2021).

Figure A7: Overall release rates under algorithmic decision rule relative to the observed release rates of judges that do not make detectable prediction mistakes.



Notes: This figure reports the overall release rate of the algorithmic decision rule that fully automates decisions against the judge's observed release rates among among judges whose choices were consistent with expected utility maximization behavior at accurate beliefs. These decisions 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 $U^*(0,0)$ (i.e., an unnecessary detention). 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 6.2 for further details. Source: Rambachan and Ludwig (2021).

Table A1: Estimated lower bound on the fraction of judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs about any pretrial misconduct risk given defendant characteristics.

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

Notes: This table summarizes the results for testing whether the release decisions of each judge in the top 25 are consistent with expected utility maximization behavior at strict preference utility functions $U(c, y^*; w)$ that (i) do not depend on any characteristics, (ii) depend on the defendant's race, (iii) depend on both the defendant's race and age, and (iv) depend on both the defendant's race and whether the defendant was charged with a felony offense. The outcome Y^* is whether the defendant would commit any pretrial misconduct (i.e., either fail to appear in court or be re-arrested for a new crime) upon release. Bounds on the any pretrial misconduct rate among detained defendants are constructed using the judge leniency instrument (see Section 5.3.2). I first construct the unadjusted rejection rate by testing whether the pretrial release decisions of each judge in the top 25 are consistent with the moment inequalities in Corollary 3.4 at the 5% level using the conditional least-favorable hybrid test using the same procedure described in Section 5.4. The adjusted rejection rate reports the fraction of rejections after correcting for multiple hypothesis testing using the Holm-Bonferroni step down procedure, which controls the family-wise error rate at the 5% level. See Section 5.4.1 for further discussion of this table. Source: Rambachan and Ludwig (2021).

B User's Guide to Identifying Prediction Mistakes in Screening Decisions

This section provides a step-by-step guide on how the identification results in Sections 3-4 for a screening decision with a binary outcome may be applied by empirical researchers.

Suppose we observe a decision maker making many decisions, and for each decision we observe the characteristics of the decision (W, X), decision maker's choice $C \in \{0, 1\}$, and the outcome $Y := C \cdot Y^*$, where $Y^* \in \{0, 1\}$ is the latent outcome of interest. We observe the dataset $\{(W_i, X_i, C_i, Y_i)\}_{i=1}^n$, which is an i.i.d. sample from the joint distribution $(W, X, C, Y) \sim P$. As discussed in Section 2.2, pretrial release decisions, medical testing and diagnostic decisions, and hiring decisions are all examples of a screening decision. The basic algorithm for testing whether the decision maker makes systematic prediction mistakes about the outcome Y^* based on the characteristics (W, X) is:

Step 1: Specify which characteristics W directly affect preferences $U(c, y^*; w)$ As discussed in Section 2.3 and Section 3.1, this is the key restriction on behavior imposed by the expected utility maximization model. Researchers may motivate this choice in two ways. First, in a particular setting, there may be common assumptions about preferences in existing empirical research, and so the researcher may directly appeal to these established modelling choices to guide this exclusion restriction. Second, the exclusion restriction may be chosen to summarize various social or legal restrictions on what characteristics ought not to directly affect preferences. I recommend that researchers report a sensitivity analysis that examines how their conclusions change under alternative assumptions about which characteristics W directly affect preferences. This explores the extent to which conclusions about systematic prediction mistakes are robust to alternative assumptions on preferences.

Step 2: Construct partitions $D_w(X)$ of the excluded characteristics X For each value of the characteristics $w \in \mathcal{W}$, construct a partition of the remaining excluded characteristics $D_w \colon \mathcal{X} \to \{1, \ldots, N_d\}$ as discussed in Section 3.3. This serves as a useful dimension-reduction step for testing whether the decision maker makes systematic prediction mistakes and conducting inference on how biased are the decision maker's predictions. The researcher may approach constructing such partitions in two ways. First, there may be data on held-out decisions (e.g., decisions made by other decision makers), and so the partition may be constructed using supervised machine learning based prediction methods to predict the outcome on the held out decisions. In this case, estimate a prediction function $\hat{f} \colon \mathcal{W} \times \mathcal{X} \to [0, 1]$ on a set of held-out decisions and define $D_w(x)$ by binning the characteristics X into percentiles of predicted risk within each value $w \in \mathcal{W}$. In the empirical application, I constructed partitions by predicting whether released defendants would fail to appear in court over decisions may be all judges in New York City except for the top 25 judges (see Section 5.3.1). Second, there may be an existing integer-valued risk score that can be used to construct the partitions and the researcher may simply choose the partition $D_w(x)$ to be level-sets associated with this existing risk score.

Step 3: Estimate the observable choice-dependent outcome probabilities Given the partition $D_w \colon \mathcal{X} \to \{1, \ldots, N_d\}$ for each $w \in \mathcal{W}$, estimate the observable choice-dependent outcome

probabilities at each cell (w, d)

$$\hat{P}_{Y^*}(1 \mid 1, w, d) := \frac{n^{-1}\sum_{i=1}^{n} C_i Y_i 1\{W_i = w, D_w(X_i) = d\}}{n^{-1}\sum_{i=1}^{n} C_i 1\{W_i = w, D_w(X_i) = d\}}.$$

Step 4: Estimate bounds on unobservable choice-dependent outcome probabilities Construct an upper bound on the unobservable choice-dependent outcome probabilities $\overline{P}_{Y^*}(1 \mid 0, w, d)$ at each cell $w \in \mathcal{W}$, $d \in \{1, \ldots, N_d\}$. This may be done in several ways.

First, as discussed in Section 3.2, there may be an instrument $Z \in \mathcal{Z}$ that is randomly assigned and generates random variation in the decision maker's choices. In this case, we can construct an upper bound on $P_{Y^*}(1 \mid 0, w, d, z)$ at each value $\tilde{z} \in \mathcal{Z}$ of the form $\frac{\pi_0(w,d,\tilde{z}) + P_{C,Y^*}(1,1|w,d,\tilde{z})}{\pi_0(w,d,z)}$. If the instrument is randomly assigned (i.e., satisfies $(Y^*, W, X) \perp\!\!\!\perp Z$), then this could be estimated directly using

$$\hat{\overline{P}}_{Y^*,\tilde{z}}(1 \mid 0, w, d, z) = \frac{\hat{\pi}_0(w, d, \tilde{z}) + \hat{P}_{C,Y^*}(1, 1 \mid w, d, \tilde{z})}{\hat{\pi}_0(w, d, z)} - \frac{\hat{P}_{C,Y^*}(1, 1 \mid w, d, z)}{\hat{\pi}_0(w, d, z)},$$

where the sample analogues are $\hat{\pi}_0(w, d, z) := \frac{n^{-1}\sum_{i=1}^{n}\sum_{i=1}^{n}(1-C_i)1\{W_i=w, D_w(X_i=d)\}}{n^{-1}\sum_{i=1}^{n}\sum_{i=1}^{n}1\{W_i=w, D_w(X_i=d)\}}$ and $\hat{P}_{C,Y^*}(1, 1 \mid w, d, z) := \frac{n^{-1}\sum_{i=1}^{n} C_i Y_i 1\{W_i=w, D_w(X_i)=d\}}{n^{-1}\sum_{i=1}^{n} 1\{W_i=w, D_w(X_i)=d\}}$. If the instrument is quasi-randomly assigned and satisfies $(Y^*, W, X) \perp\!\!\!\perp Z \mid T$, then apply the identification results in Appendix D.1. In the empirical application, I used the quasi-random assignment of judges to cases to construct bounds on the unobservable failure to appear rate among detained defendants (see Section 5.3.2).

Second, as mentioned in Remark 3.3, researchers may introduce additional assumptions that bound the unobservable choice-dependent outcome probabilities using the observable choice-dependent outcome probabilities. I refer to this as "direct imputation." In direct imputation, the researcher specifies $\kappa_{w,d} \geq 0$ for each cell $w \in \mathcal{W}$, $d \in \{1, \ldots, N_d\}$ and assumes that $P_{Y^*}(1 \mid 0, w, d) \leq (1 + \kappa_{w,d}) P_{Y^*}(1 \mid 1, w, d)$. See Supplement G for further details. I recommend that researchers report a sensitivity analysis on their conclusions based on the choices of $\kappa_{w,d} \geq 0$. I illustrate such a sensitivity analysis for direct imputation in Supplement I for the New York City pretrial release setting.

Finally, as mentioned in Remark 3.3, researchers may also observe an additional proxy outcome that does not suffer from the missing data problem, and can therefore be used to construct bounds on the unobservable choice-dependent outcome probabilities provided the researcher introduces bounds on the joint distribution of the proxy outcome and the latent outcome. See Supplement G for further details.

Step 5: Testing whether the decision maker makes systematic prediction mistakes Testing whether the decision maker makes systematic prediction mistakes given a choice of directly payoff-relevant characteristics, a partition of the excluded characteristics and bounds on the unobservable choice-dependent outcome probabilities is equivalent to testing whether the moment inequalities in Corollary 3.4 are satisfied. Testing whether these moment inequalities are satisfied tests the null hypothesis that the decision maker's choices are consistent with expected utility maximization behavior at preferences that satisfy the researcher's conjectured exclusion restriction. Researchers may pick their preferred moment inequality testing procedure from the econometrics literature.

See, for example, the reviews in Canay and Shaikh (2017) and Molinari (2020). In the empirical application, I use the conditional least-favorable hybrid test developed in Andrews, Roth and Pakes (2019) since it is computationally fast given estimates of the moments and the variance-covariance matrix and has desirable power properties.

Step 6: Conducting Inference on how biased are the decision maker's predictions To conduct inference on how biased are the decision maker's predictions between cells $w \in \mathcal{W}, d \in \{1, \ldots, N_d\}$ and $w \in \mathcal{W}, d' \in \{1, \ldots, N_d\}$, first construct a joint confidence set for the decision maker's reweighed utility thresholds $\tau(w,d), \tau(w,d')$ at cells $(w,d), (w,d')$ based on the moment inequalities in (9). This can be done through test-inversion: for a grid of possible values for the reweighted thresholds, test the null hypothesis that the moment inequalities in (9) are satisfied at each point in the grid and collect together all points at which we fail to reject the null hypothesis. Second, for each value in the joint confidence set, construct the ratio in (10). This provides a confidence set for the decision maker's implied prediction mistake between cells (w,d) and (w,d'). If this confidence set for the implied prediction mistake lies everywhere below one, then the decision maker's beliefs about the latent outcome given the characteristics are underreacting to predictable variation in the latent outcome. Analogously, if this confidence set for the implied prediction mistake lies everywhere above one, then the decision maker's beliefs are overreacting. See Section 4.2 for further discussion on the interpretation of this implied prediction mistake.

C Additional Results for the Expected Utility Maximization Model

In this section of the appendix, I provide additional results for the expected utility maximization model that are mentioned in the main text.

C.1 Characterization of Expected Utility Maximization in Treatment Decisions

In Section 3, I analyzed the testable implications of expected utility maximization behavior at accurate beliefs in screening decisions with a binary outcome. I now show that these identification results extend to treatment decisions with a multi-valued outcome for particular classes of utility functions $\mathcal{U}$. These extensions also apply to screening decisions with a multi-valued outcome.

C.1.1 Linear Utility

First, I analyze the conditions under which the decision maker's choices are consistent with expected utility maximization at a linear utility function of the form $U(c, \vec{y}; w) = \beta(w)y - \lambda(w)c$, where $\mathcal{Y} \in \mathbb{R}$ and $\beta(w) > 0, \lambda(w) > 0$ for all $w \in \mathcal{W}$. This is an extended Roy model in which the benefit function only depends on the realized outcome linearly.[41]

As notation, let $\mu_{Y_1 - Y_0}(c, w, x) := \mathbb{E}[Y_1 - Y_0 \mid C = c, W = w, X = x]$. Define $\mathcal{X}^0(w) := \{x \in \mathcal{X} \colon \pi_0(w,x) > 0\}$ and $\mathcal{X}^1(w) := \{x \in \mathcal{X} \colon \pi_1(w,x) > 0\}$.

[41] Henry, Meango and Mourifie (2020) studies an extended Roy model under the assumption that utility function satisfies $U(0, \vec{y}; w) = y_0, U(1, \vec{y}; w) = Y_1 - \lambda(Y_1)$ for some function $\lambda(\cdot)$. The authors derive testable restrictions on behavior under this extended Roy model provided the researcher observes a stochastically monotone instrumental variable.

Theorem C.1. *Consider a treatment decision. The decision maker's choices are consistent with expected utility maximization at some utility function $U(c, \vec{y}; w) = \beta(w)Y - \lambda(w)C$ if and only if, for all $w \in \mathcal{W}$,*

$$\max_{x \in \mathcal{X}^0(w)} \underline{\mu}_{Y_1 - Y_0}(0, w, x) \leq \min_{x \in \mathcal{X}^1(w)} \overline{\mu}_{Y_1 - Y_0}(1, w, x),$$

where

$$\underline{\mu}_{Y_1 - Y_0} = \min\{\mu_{Y_1 - Y_0}(0, w, x) \colon \widetilde{P}_{\vec{Y}}(\cdot \mid 0, w, x) \in \mathcal{B}_{0,w,x}\},$$

$$\overline{\mu}_{Y_1 - Y_0}(1, w, x) = \max\{\mu_{Y_1 - Y_0}(1, w, x) \colon \widetilde{P}_{\vec{Y}}(\cdot \mid 1, w, x) \in \mathcal{B}_{1,w,x}\}.$$

Proof. This result follows immediately from applying the inequalities in Theorem 2.1. Over $(w, x) \in \mathcal{W} \times \mathcal{X}$ such that $\pi_1(w, x) > 0$, $\frac{\lambda(w)}{\beta(w)} \leq \mu_{Y_1 - Y_0}(1, w, x)$ must be satisfied: Analogously, over $(w, x) \in \mathcal{W} \times \mathcal{X}$ such that $\pi_0(w, x) > 0$, $\frac{\lambda(w)}{\beta(w)} \geq \mu_{Y_1 - Y_0}(0, w, x)$ must be satisfied. The result is then immediate. □

In a treatment decision with a binary outcome, Theorem C.1 immediately implies negative results about the testability of expected utility maximization behavior that are analogous to those stated in the main text for a screening decision with binary outcomes. I state these negative results as a corollary. Define

$$\underline{P}_{\vec{Y}}(0, 1 \mid 0, w, x) - \underline{P}_{\vec{Y}}(1, 0 \mid 0, w, x) =$$

$$\min \left\{ \widetilde{P}_{\vec{Y}}(0, 1 \mid 0, w, x) - \widetilde{P}_{\vec{Y}}(1, 0 \mid 0, w, x) \colon \widetilde{P}_{\vec{Y}}(\cdot \mid 0, w, x) \in \mathcal{B}_{0,w,x} \right\}$$

and

$$\overline{P}_{\vec{Y}}(0, 1 \mid 1, w, x) - \overline{P}_{\vec{Y}}(1, 0 \mid 1, w, x) =$$

$$\max \left\{ \widetilde{P}_{\vec{Y}}(0, 1 \mid 1, w, x) - \widetilde{P}_{\vec{Y}}(1, 0 \mid 1, w, x) \colon \widetilde{P}_{\vec{Y}}(\cdot \mid 1, w, x) \in \mathcal{B}_{1,w,x} \right\}.$$

Corollary C.1. *Consider a treatment decision with a binary outcome. The decision maker's choices are consistent with expected utility maximization at some utility function $U(c, \vec{y}; w) = \beta(w)y - \lambda(w)c$ if either:*

(i) All characteristics affect utility (i.e., $\mathcal{X} = \emptyset$) and $\underline{P}_{\vec{Y}}(0, 1 \mid 0, w) - \underline{P}_{\vec{Y}}(1, 0 \mid 0, w) \leq \overline{P}_{\vec{Y}}(0, 1 \mid 1, w) - \overline{P}_{\vec{Y}}(1, 0 \mid 1, w)$ for all $w \in \mathcal{W}$.

(ii) The researcher's bounds on the choice-dependent potential outcome probabilities are uninformative, meaning that for both $c \in \{0, 1\}$, $\mathcal{B}_{c,w,x}$ is the set of all $\widetilde{P}_{\vec{Y}}(\cdot \mid c, w, x)$ satisfying $\sum_{y_{\tilde{c}} \in \mathcal{Y}} \widetilde{P}_{\vec{Y}}(\cdot, y_{\tilde{c}} \mid c, w, x) = P_{Y_c}(\cdot \mid c, w, x)$ for all $\widetilde{P}_{\vec{Y}}(\cdot \mid c, w, x) \in \mathcal{B}_{c,w,x}$.

Proof. Case (i) follows immediately from Theorem C.1. Case (ii) follows since under uninformative bounds on the missing data, $\overline{P}_{\vec{Y}}(0, 1 \mid 1, w, x) - \overline{P}_{\vec{Y}}(1, 0 \mid 1, w, x) = P_{Y_1}(1 \mid 1, w, x)$ and $\underline{P}_{\vec{Y}}(0, 1 \mid 0, w, x) - \underline{P}_{\vec{Y}}(1, 0 \mid 0, w, x) = -P_{Y_0}(1 \mid 0, w, x)$. □

C.1.2 Binary-Valued Utility Function

I analyze the conditions under which the decision maker's choices are consistent with expected utility maximization at a utility function that is a simple function over $\mathcal{Y}$. That is, for some known $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$, define $\tilde{Y}_c = 1\{Y_c \in \tilde{\mathcal{Y}}\}$ and $\tilde{Y} = C\tilde{Y}_1 + (1 - C)\tilde{Y}_0$. Consider the class of utility functions

of the form $u(c, \vec{y}; w) := u(c, \tilde{y}; w)$. For this class of utility function, the decision maker faces a treatment decision with a binary outcome, and so the previous analysis applies if we further assume that $u(c, \tilde{y}; w) = \beta(w)\tilde{y} - \lambda(w)c$.

C.2 ϵ_w-Approximate Expected Utility Maximization

The expected utility maximization model in the main text assumes that the decision maker exactly maximizes expected utility given their preferences and beliefs about the outcome given the characteristics and some private information. The decision maker, however, may suffer from various cognitive limitations that prevent them from exact maximization. That is, the decision maker may be boundedly rational and therefore make choices to "satisfice" rather than fully optimize (Simon, 1955, 1956). I next weaken the definition of expected utility maximization to allow the decision maker to be an ϵ_w-approximate expected utility maximizer.

Definition 8 (ϵ_w-approximate expected utility maximization)**.** The decision maker's choices are consistent with *ϵ_w-approximate expected utility maximization* if there exists a utility function $U \in \mathcal{U}$, $\epsilon_w \geq 0$ for all $w \in \mathcal{W}$, and joint distribution $(W, X, V, C, \vec{Y}) \sim Q$ satisfying:

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

$$\mathbb{E}_Q\left[U(c, \vec{Y}; W) \mid W = w, X = x, V = v\right] \geq \mathbb{E}_Q\left[U(c', \vec{Y}; W) \mid W = w, X = x, V = v\right] - \epsilon_w.$$

(ii) Information Set, (iii) Data Consistency.

That is, the decision-maker's choices are consistent with ϵ_w-approximate expected utility maximization if there choices are within $\epsilon_w \geq 0$ of being optimal. This is analogous to the proposed measure of violations of expected utility maximization in consumer optimization in Echenique, Saito and Imai (2021). Allen and Rehbeck (2020) propose a measure of whether a decision maker's choices in consumer optimization are ϵ-rationalizable by a quasi-linear utility function, whereas my interest is focused on how well the decision maker's choices are approximated by expected utility maximization in empirical treatment decisions. Apesteguia and Ballester (2015) propose a "swaps-index" to measure how much an observed preference relation violates utility maximization and expected utility maximization, which summarizes the number of choices that must be swapped in order to rationalize the data. Notice that the special case with $\epsilon_w = 0$ nests the definition of expected utility maximization given in the main text (Definition 3).

Theorem 2.1 directly extends to characterize ϵ_w-approximate expected utility maximization behavior in treatment decisions.

Theorem C.2. *The decision maker's choices are consistent with ϵ_w-expected utility maximization if and only if there exists a utility function $U \in \mathcal{U}$, $\epsilon_w \geq 0$ for all $w \in \mathcal{W}$, $\widetilde{P}_{\vec{Y}}(\cdot \mid 0, w, x) \in \mathcal{B}_{0,w,x}$ and $\widetilde{P}_{\vec{Y}}(\cdot \mid 1, w, x) \in \mathcal{B}_{1,w,x}$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$ satisfying*

$$\mathbb{E}_Q\left[U(c, \vec{Y}; W) \mid C = c, W = w, X = x\right] \geq \mathbb{E}_Q\left[U(c', \vec{Y}; W) \mid C = c, W = w, X = x\right] - \epsilon_w \tag{8}$$

for all $c \in \{0, 1\}$, $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $\pi_c(w, x) > 0$ and $c' \neq c$, where the joint distribution $(W, X, C, \vec{Y}) \sim Q$ is given by $Q(w, x, c, \vec{y}) = \widetilde{P}_{\vec{Y}}(\vec{y} \mid c, w, x)P(c, w, x)$.

Proof. The proof follows the same argument as the proof of Theorem 2.1. □

In words, the decision maker's choices are consistent with ϵ_w-approximate expected utility maximization if and only if they approximately satisfy the revealed preference inequalities derived in the main text.

The value of Theorem C.2 comes in applying the approximate revealed preference inequalities to analyze particular decision problems. I next use this result to characterize whether the decision maker's choices are consistent with approximate expected utility maximization at strict preferences in a screening decision with a binary outcome.

Theorem C.3. *Consider a screening decision with a binary outcome. Assume $P_{Y^*}(1 \mid 1, w, x) < 1$ for all $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $\pi_1(w, x) > 0$. The decision maker's choices are consistent with ϵ_w-approximate expected utility maximization at some strict preference utility function if and only if, for all $w \in \mathcal{W}$, there exists $\tilde{\epsilon}_w \geq 0$ satisfying*

$$\max_{x \in \mathcal{X}^1(w)} P_{Y^*}(1 \mid 1, w, x) - \tilde{\epsilon}_w \leq \min_{x \in \mathcal{X}^0(w)} \overline{P}_{Y^*}(1 \mid 0, w, x) + \tilde{\epsilon}_w.$$

where $\mathcal{X}^1(w) := \{x \in \mathcal{X} \colon \pi_1(w, x) > 0\}$ and $\mathcal{X}^0(w) := \{x \in \mathcal{X} \colon \pi_0(w, x) > 0\}$.

Proof. The proof follows the same argument as the Proof of Theorem 3.1, which I provide for completeness. For all $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $\pi_1(w, x) > 0$, Theorem C.2 requires that $P_{Y^*}(1 \mid 1, w, x) \leq \frac{U(0,0;w)}{U(0,0;w)+U(1,1;w)} + \frac{\epsilon_w}{-U(0,0;w)-U(1,1;w)}$. Analogously, or all $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $\pi_0(w, x) > 0$, Theorem C.2 requires that $\frac{U(0,0;w)}{U(0,0;w)+U(1,1;w)} - \frac{\epsilon_w}{-U(0,0;w)-U(1,1;w)} \leq P_{Y^*}(1 \mid 0, w, x)$. The result is immediate after defining $\tilde{\epsilon}_w \geq 0$ as $\frac{\epsilon_w}{-U(0,0;w)-U(1,1;w)}$. □

That is, the decision maker's choices are consistent with ϵ_w-approximate expected utility maximization if and only if the decision maker is acting as if she applies an approximate, incomplete threshold rule in selecting her choices. This means that the observed choice-dependent outcome probabilities and bounds on the unobservable choice-dependent outcome probabilities must satisfy a relaxation of the inequalities given in Theorem 3.1. As shown in the proof, the relaxation satisfies $\tilde{\epsilon}_w = \frac{\epsilon_w}{-U(0,0;w)-U(1,1;w)}$. Notice that as ϵ_w grows large for each $w \in \mathcal{W}$, the decision maker's choices are always rationalizable under ϵ_w-approximate expected utility maximization. Therefore, searching for the minimal value $\tilde{\epsilon}_w \geq 0$ such that the inequalities are satisfied provides a simple summary measure of how "far from optimal" are the decision maker's choices. This minimal value is given by

$$\tilde{\epsilon}_w = \left(\max_{x \in \mathcal{X}^1(w)} P_{Y^*}(1 \mid 1, w, x) - \min_{x \in \mathcal{X}^0(w)} \overline{P}_{Y^*}(1 \mid 0, w, x) \right)_+,$$

where $(A)_+ = \max\{A, 0\}$. Alternatively, the minimal value can also be characterized as the smallest $\tilde{\epsilon}_w \geq 0$ that satisfies the following system of moment inequalities

$$P_{Y^*}(1 \mid 1, w, x) - \overline{P}_{Y^*}(1 \mid 0, w, x') - 2\tilde{\epsilon}_w \leq 0$$

for all $w \in \mathcal{W}$, $(x, x') \in \mathcal{X}$.

Finally, the implied revealed preference inequalities over partitions of the excluded characteristics $x \in \mathcal{X}$ given in Corollary 3.4 can analogously be extended. In particular, for partitions

$D_w \colon \mathcal{X} \to \{1, \ldots, N_d\}$, if the decision maker's choices are consistent with ϵ_w-approximate expected utility maximization, then there exists $\tilde{\epsilon}_w \geq 0$ satisfying

$$\max_{x \in \mathcal{X}^1(w)} P_{Y^*}(1 \mid 1, w, d) - \tilde{\epsilon}_w \leq \min_{x \in \mathcal{X}^0(w)} \overline{P}_{Y^*}(1 \mid 0, w, d) + \tilde{\epsilon}_w.$$

Therefore, researchers could characterize the implied relaxation $\tilde{\epsilon}_w \geq 0$ over the constructed partitions $D_w(X)$.

C.3 Expected Utility Maximization with Inaccurate Beliefs after Dimension Reduction

In this section, I show that the identification result for expected utility maximization behavior with inaccurate beliefs extends to coarsening the excluded characteristics. Let $D_w \colon \mathcal{X} \to \{1, \ldots, d_w\}$ be a function that partitions the observable characteristics X into level sets $\{x \in \mathcal{X} \colon D_w(x) = d\}$. The next result follows by applying iterated expectations to Lemma E.3.

Proposition C.1. *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 $w \in \mathcal{W}$, $d \in \{1, \ldots, N_d\}$, $c \in \{0, 1\}$ and $c' \neq c$,*

$$\sum_{\vec{y} \in \mathcal{Y}^{N_c}} Q_{C,\vec{Y}}(c, \vec{y} \mid w, D_w(x) = d) U(c, \vec{y}; w) \geq \sum_{\vec{y} \in \mathcal{Y}^{N_c}} Q_{C,\vec{Y}}(c, \vec{y} \mid w, D_w(x) = d) U(c, \vec{y}; w),$$

where

$$Q_{C,\vec{Y}}(c, \vec{y} \mid w, D_w(x) = d) = \left(\sum_{x \colon D_w(x) = d} P_C(c \mid \vec{y}, w, x) Q_{\vec{Y}}(\vec{y} \mid w, x) P(x \mid w) \right) / P(D_w(x) = d \mid w),$$

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

Provided that $P_{C,\vec{Y}}(c, \vec{y} \mid w, x) > 0$ for all $(c, \vec{y}) \in \mathcal{C} \times \mathcal{Y}^{N_c}$ and $(w, x) \in \mathcal{W} \times \mathcal{X}$, Proposition C.1 can be recast as checking whether there exists non-negative weights $\omega(c, \vec{y}; w, d) \geq 0$ satisfying

$$\sum_{\vec{y} \in \mathcal{Y}^{N_c}} \omega(c, \vec{y}; w, d) P_{C,\vec{Y}}(c, \vec{y} \mid w, D_w(x) = d) U(c, \vec{y}; w) \geq$$

$$\sum_{\vec{y} \in \mathcal{Y}^{N_c}} \omega(c, \vec{y}; w, d) P_{C,\vec{Y}}(c, \vec{y} \mid w, D_w(x) = d) U(c, \vec{y}; w)$$

and $\mathbb{E}_P \left[\omega(C, \vec{Y}; W, D_w(X)) \mid W = w, D_w(x) = d \right] = 1$.

I next apply this result in a screening decision with a binary outcome. In this special case, this result may be applied to derive bounds on the decision maker's reweighed utility threshold through

$$P_{Y^*}(1 \mid 1, w, d) \leq \frac{\omega(0, 0; w, d) U(0, 0; w)}{\omega(0, 0; w, d) U(0, 0; w) + \omega(1, 1; w, d) U(1, 1; w)} \leq \overline{P}_{Y^*}(1 \mid 0, w, d), \quad (9)$$

where $P_{Y^*}(y^* \mid c, w, d) := P(Y^* = y^* \mid C = c, W = w, D_w(X) = d)$. Next, define $M = 1\{C = 0, Y^* = 0\} + 1\{C = 1, Y^* = 1\}$, $\tau(w, d) = \frac{\omega(0,0;w,d)U(0,0;w)}{\omega(0,0;w,d)U(0,0;w)+\omega(1,1;w,d)U(1,1;w)}$. Examining $w \in \mathcal{W}$, $d, d' \in \{1, \ldots, N_d\}$, we arrive at

$$\frac{(1-\tau(w,d))/\tau(w,d)}{(1-\tau(w,d'))/\tau(w,d')} = \frac{\frac{Q(C=1,Y^*=1|M=1,w,d)/Q(C=0,Y^*=0|M=1,w,d)}{Q(C=1,Y^*=1|M=1,w,d')/Q(C=0,Y^*=0|M=1,w,d')}}{\frac{P(C=1,Y^*=1|M=1,w,d)/P(C=0,Y^*=0|M=1,w,d)}{P(C=1,Y^*=1|M=1,w,d')/P(C=0,Y^*=0|M=1,w,d')}}. \tag{10}$$

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 her own "ex-post mistakes." 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 are under-reacting to variation in risk across the cells (w, d) and (w, d'). If these bounds lie everywhere above one, then the decision maker's beliefs are over-reacting.

C.4 The Policymaker's First-Best Decision Rule

Consider a policymaker with social welfare function $U^*(0,0) < 0, U^*(1,1) < 0$ as in Section 6.1. I construct an algorithmic decision rule based on analyzing how the policymaker would make choices herself in the binary screening decision. Rambachan et al. (2021) refer to this as the "first-best problem" in their analysis of algorithmic decision rules.

Due to the missing data problem, the conditional probability of $Y^* = 1$ given the characteristics is partially identified and I assume the policymaker adopts a max-min evaluation criterion to evaluate decision rules. Let $p^*(w, x) \in [0, 1]$ denote the probability the policymaker selects $C = 1$ given $W = w, X = x$. At each $(w, x) \in \mathcal{W} \times \mathcal{X}$, the policymaker chooses $p^*(w, x)$ to maximize

$$\min_{\widetilde{P}_{Y^*}(1|w,x)} p^*(w,x)\widetilde{P}_{Y^*}(1 \mid w,x)U^*(1,1) + (1-p^*(w,x))(1-\widetilde{P}_{Y^*}(1 \mid w,x))U^*(0,0)$$
$$\text{s.t. } \underline{P}_{Y^*}(1 \mid w,x) \leq \widetilde{P}_{Y^*}(1 \mid w,x) \leq \overline{P}_{Y^*}(1 \mid w,x).$$

Proposition C.2. *Consider a binary screening decision and a policymaker with social welfare function $U^*(0,0) < 0$, $U^*(1,1) < 0$, who chooses $p^*(w,x) \in [0,1]$ to maximize her worst-case expected utility. Defining $\tau^*(U^*) := \frac{U^*(0,0)}{U^*(0,0)+U^*(1,1)}$, her max-min decision rule is*

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

Proof. To show this result, I consider cases.

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

$$P(Y^* = 1 \mid W = w, X = x)U^*(1,1) \geq P(Y^* = 0 \mid W = w, X = x)U^*(0,0)$$

for all $P(Y^* = 1 \mid W = w, X = x)$ satisfying $\underline{P}(Y^* = 1 \mid W = w, X = x) \leq P(Y^* = 1 \mid W =$

$w, X = x) \leq \overline{P}(Y^* = 1 \mid W = w, X = x)$. Therefore, it is optimal to set $p^*(w, x) = 1$ in this case.

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

$$P(Y^* = 1 \mid W = w, X = x)U^*(1, 1) \leq P(Y^* = 0 \mid W = w, X = x)U^*(0, 0)$$

for all $P(Y^* = 1 \mid W = w, X = x)$ satisfying $\underline{P}(Y^* = 1 \mid W = w, X = x) \leq P(Y^* = 1 \mid W = w, X = x) \leq \overline{P}(Y^* = 1 \mid W = w, X = x)$. Therefore, it is optimal to set $p^*(w, x) = 0$ in this case.

Case 3: Suppose $\underline{P}(Y^* = 1 \mid W = w, X = x) < \tau^* < \overline{P}(Y^* = 1 \mid W = w, X = x)$. Begin by noticing that $p^*(w, x) = \tau^*$ delivers constant expected payoffs for all $P(Y^* = 1 \mid W = w, X = x)$ satisfying $\underline{P}(Y^* = 1 \mid W = w, X = x) \leq P(Y^* = 1 \mid W = w, X = x) \leq \overline{P}(Y^* = 1 \mid W = w, X = x)$. As a function of $P(Y^* = 1 \mid W = w, X = x)$ and $p^*(w, x)$, expected social welfare equals

$$p^*(w, x)P(Y^* = 1 \mid W = w, X = x)U^*(1, 1)+(1-p^*(w, x))P(Y^* = 0 \mid W = w, X = x)U^*(0, 0).$$

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

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

$$p^*(w, x)\underline{P}(Y^* = 1 \mid W = w, X = x)U^*(1, 1)+(1-p^*(w, x))(1-\underline{P}(Y^* = 1 \mid W = w, X = x))U^*(0, 0) <$$

$$\tau^*\underline{P}(Y^* = 1 \mid W = w, X = x)U^*(1, 1) + (1 - \tau^*)(1 - \underline{P}(Y^* = 1 \mid W = w, X = x))U^*(0, 0) =$$

$$\frac{U^*(0, 0)U^*(1, 1)}{U^*(0, 0) + U^*(1, 1)}.$$

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

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

$$p^*(w, x)\overline{P}(Y^* = 1 \mid W = w, X = x)U^*(1, 1)+(1-p^*(w, x))(1-\overline{P}(Y^* = 1 \mid W = w, X = x))U^*(0, 0) <$$

$$\tau^*\overline{P}(Y^* = 1 \mid W = w, X = x)U^*(1, 1) + (1 - \tau^*)(1 - \overline{P}(Y^* = 1 \mid W = w, X = x))U^*(0, 0) =$$

$$\frac{U^*(0, 0)U^*(1, 1)}{U^*(0, 0) + U^*(1, 1)}.$$

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

The policymaker makes choices based on a threshold rule, where the threshold τ^* depends on the relative costs to ex-post errors assigned by the social welfare function. If the upper bound on the probability of $Y^* = 1$ 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 $P(Y = 1 \mid W = w, X = x)$ contains the threshold τ^*, the policymaker randomizes her decision and selects $C = 1$ with probability exactly equal to τ^*.

In my empirical analysis in Section 6.2, I evaluate the choices of judges against this first-best decision rule applied to each cell of payoff relevant characteristics W and each decile of predicted risk $D_w(X)$. The bounds on the probability of $Y^* = 1$ conditional on the characteristics is constructed using the quasi-random assignment of judges as discussed in Section 5.3.2, and the threshold τ^* varies as the social welfare function $U^*(0,0), U^*(1,1)$ varies.

D Additional Results for the Econometric Framework

In this section of the appendix, I provide additional results that are useful for implementing the econometric framework on testing for and characterizing systematic prediction mistakes.

D.1 Constructing Bounds on the Missing Data through a Quasi-Random Instrument

In this section, I modify Assumption 3.1 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 $(W, X, T, Z, C, Y^*) \sim P$ satisfies

$$(W, X, Y^*) \perp\!\!\!\perp Z \mid T \tag{11}$$

and $P(w,x,t,z) > 0$ for all $(w,x,t,z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{T} \times \mathcal{Z}$. This is useful as in my empirical application to pretrial release decisions in New York City, bail judges are quasi-randomly assigned to cases within a court-by-time cell. See Section 5 for details.

Under (11), researchers can derive bounds on the unobservable choice-dependent outcome probabilities in a screening decision with a binary latent outcome. By iterated expectations,

$$P_{Y^*}(1 \mid w,x,z) = \sum_{t \in \mathcal{T}} P_{Y^*}(1 \mid w,x,z,t) P(t \mid w,x,z) = \sum_{t \in \mathcal{T}} P_{Y^*}(1 \mid w,x,\tilde{z},t) P(t \mid w,x,z),$$

where the last equality follows by quasi-random assignment. Furthermore, for each value of $t \in \mathcal{T}$ and $z \in \mathcal{Z}$, $P_{Y^*}(1|w,x,z,t)$ is bounded by

$$P_{C,Y^*}(1,1 \mid w,x,z,t) \leq P_{Y^*}(1 \mid w,x,z,t) \leq \pi_0(w,x,z,t) + P_{C,Y^*}(1,1 \mid w,x,z,t).$$

Therefore, for a given $z \in \mathcal{Z}$, valid lower and upper bounds on $P_{Y^*}(1 \mid w,x,z)$ are given by

$$\begin{aligned}
&\mathbb{E}\left[P_{C,Y^*}(1,1 \mid W,X,\tilde{z},T) \mid W = w, X = x, Z = z\right] \leq P_{Y^*}(1 \mid w,x,z), \\
&P_{Y^*}(1 \mid w,x,z) \leq \mathbb{E}\left[P_{C,Y^*}(1,1 \mid W,X,\tilde{z},T) + \pi_0(W,X,\tilde{z},T) \mid W = w, X = x, Z = z\right]
\end{aligned}$$

for any $\tilde{z} \in \mathcal{Z}$. Since $P_{C,Y^*}(1,1 \mid w,x,z)$ is observed, this naturally implies bounds on $P_{C,Y^*}(0,1 \mid w,x,z)$. This in turn gives a bound on $P_{Y^*}(1 \mid 0,w,x,z)$ since $\pi_c(w,x,z)$ is also observed (assuming $\pi_c(w,x,z) > 0$).

D.2 Testing Expected Utility Maximization Behavior in Treatment Decisions

In this section, I extend the econometric framework for analyzing screening decisions in Section 3 to treatment decisions. First, I discuss how the researcher may construct bounds on the unobservable choice-dependent potential outcome probabilities using an instrument. Second, I discuss how testing the revealed preference inequalities in these settings reduces to testing many moment inequalities with nuisance parameters that enter linearly.

D.2.1 Constructing Bounds with an Instrument in Treatment Decisions

As in the main text, let $Z \in \mathcal{Z}$ be a finite support instrument. Assume that the joint distribution $(W,X,Z,C,\vec{Y}) \sim P$ satisfies $(W,X,\vec{Y}) \perp\!\!\!\perp Z$ and $P(W=w,X=x,Z=z) > 0$ for all $(w,x,z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Z}$. In this case, the conditional joint distribution $(C,\vec{Y}) \mid W,X,Z$ is partially identified and the next result provides bounds on this quantity.

Proposition D.1. *Consider any* $(w,x,z) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Z}$. *If* $\widetilde{P}_{C,\vec{Y}}(\cdot,\cdot \mid w,x,z) \in \mathcal{H}_P(P_{C,\vec{Y}}(\cdot,\cdot \mid w,x,z))$, *then it satisfies*

(i) For all $y \in \mathcal{Y}$,

$$\sum_{y_0 \in \mathcal{Y}} \widetilde{P}_{C,\vec{Y}}(1,y_0,y \mid w,x,z) = P_{C,Y_1}(1,y \mid w,x,z)$$
$$\sum_{y_1 \in \mathcal{Y}} \widetilde{P}_{C,\vec{Y}}(0,y,y_1 \mid w,x,z) = P_{C,Y_0}(0,y \mid w,x,z).$$

(ii) For all $\tilde{z} \in \mathcal{Z}$, $\vec{y} \in \mathcal{Y} \times \mathcal{Y}$,

$$0 \leq \widetilde{P}_{C,\vec{Y}}(1,\vec{y} \mid w,x,z) + \widetilde{P}_{C,\vec{Y}}(0,\vec{y} \mid w,x,z) \leq P_{C,Y_1}(1,y_1 \mid w,x,\tilde{z}) + P_{C,Y_0}(0,y_0 \mid w,x,\tilde{z}).$$

Proof. Consider a particular value $(w,x,z) \in \mathcal{W} \times \mathcal{Z}$. Notice that $P_{\vec{Y}}(\cdot \mid w,x,z) = P_{\vec{Y}}(\cdot \mid w,x,\tilde{z})$ by random assignment of the instrument. Furthermore, notice that $P_{\vec{Y}}(\vec{y} \mid w,x,z) = P_{C,\vec{Y}}(0,\vec{y} \mid w,x,z) + P_{C,\vec{Y}}(1,\vec{y} \mid w,x,z)$, where

$$0 \leq P_{C,\vec{Y}}(0,\vec{y} \mid w,x,z) \leq P_{C,Y_0}(0,y_0 \mid w,x,z),$$
$$0 \leq P_{C,\vec{Y}}(1,\vec{y} \mid w,x,z) \leq P_{C,Y_1}(0,y_1 \mid w,x,z).$$

Therefore, we observe that $P_{\vec{Y}}(\vec{y} \mid w,x,z)$ must be bounded below by 0 (trivially) and bounded above by $P_{C,Y_0}(0,y_0 \mid w,x,z) + P_{C,Y_1}(0,y_1 \mid w,x,z)$ for each $z \in \mathcal{Z}$. The result is then immediate by also requiring data consistency. □

These simple bounds are non-sharp, but they can be tightened by using Arstein's Theorem. Results in Russell (2019) and Kitagawa (2020) characterize the sharp identified set of potential outcomes in treatment assignment problems and imply sharp bounds on $P_{\vec{Y}}(\cdot \mid w,x,z)$. We can

replace the non-sharp bounds in (ii) in Proposition D.1 with these sharp bounds to obtain tight bounds on the conditional joint distribution $(C, \vec{Y}) \mid W, X, Z$. The number of inequalities in these sharp bounds grow exponentially in the support of the potential outcomes and equals $2^{\mathcal{Y}\times\mathcal{Y}}$.

D.2.2 Testing for Prediction Mistakes Reduces to Testing Many Moment Inequalities with Linear Nuisance Parameters

Suppose the researcher wishes to test whether the decision maker's choices are consistent with expected utility maximization behavior at some utility function $U \in \mathcal{U}$ in a general screening decision, where recall that $\mathcal{U}$ is the set of feasible utility functions specified by the researcher. Denote this null hypothesis by $H_0(\mathcal{U})$ As a stepping stone, I provide a reduction to show how the researcher may test whether the decision maker's choices are consistent with expected utility maximization behavior at a particular utility function $U \in \mathcal{U}$. Denote this particular null hypothesis as $H_0(U)$. As discussed in Bugni, Canay and Shi (2015), the researcher may construct a conservative test of $H_0(\mathcal{U})$ by constructing a confidence interval for the identified set of utility functions through test inversion on $H_0(U)$ and checking whether this confidence interval is empty.

With this in mind, consider a fixed utility function $U \in \mathcal{U}$ and suppose the researcher constructs bounds on the unobservable choice-dependent outcome probabilities using a randomly assigned instrument. Testing $H_0(U)$ is equivalent to testing a possibly high-dimensional set of moment inequalities with linear nuisance parameters. I will prove this result for the non-sharp bounds stated in Proposition D.1 but the result extends to using sharp bounds based on Arstein's Inequality.

Proposition D.2. *Let $N_y := |\mathcal{Y}|$. Assume there is a randomly assigned instrument. The decision maker's choices at $z \in \mathcal{Z}$ are consistent with expected utility maximization behavior at utility function $U\colon \{0,1\} \times \mathcal{Y} \times \mathcal{Y} \times \mathcal{W} \to \mathbb{R}$ if there exists $\tilde{\delta} \in \mathbb{R}^{2d_w d_x N_y (N_y - 1)}$ satisfying*

$$\tilde{A}(U)_{(\cdot,1:m)}\mu(P) + \tilde{A}(U)_{(\cdot,-(1:m))}\tilde{\delta} \leq b,$$

where $\tilde{A}(U)$ is a matrix of known constants that depend on the specified utility function U, b is a vector of known constants, $\mu(P)$ is a $m := 2d_w d_x N_y + d_w d_x N_y^2 (N_z - 1)$ dimensional vector that collects together the observable moments and bounds based on the instrument.[42]

Proof. Applying the non-sharp bounds in Proposition D.1, I begin by restating Lemma E.1 in terms of the nuisance parameters $\widetilde{P}_{C,\vec{Y}}(\cdot \mid w, x, z) \in \Delta(\mathcal{C} \times \mathcal{Y} \times \mathcal{Y})$ with entries

$$\widetilde{P}_{C,\vec{Y}}(\cdot \mid w, x, z) = \begin{pmatrix} P_{C,\vec{Y}}(c_1, \vec{y}_1 \mid w, x, z) \\ \vdots \\ P_{C,\vec{Y}}(c_1, \vec{y}_{N_y^2} \mid w, x, z) \\ \vdots \\ P_{C,\vec{Y}}(c_{N_c}, \vec{y}_1 \mid w, x, z) \\ \vdots \\ P_{C,\vec{Y}}(c_{N_c}, \vec{y}_{N_y^2} \mid w, x, z) \end{pmatrix}.$$

[42] For a matrix B, the notation $B_{(\cdot,1:m)}$ denotes the submatrix containing the first m columns of B and $B_{(\cdot,-1:m)}$ denotes the submatrix containing all columsn except the first m of B.

Lemma D.1. *The decision maker's choices at $z \in \mathcal{Z}$ are consistent with expected utility maximization behavior at utility function U if for all $(w,x) \in \mathcal{W} \times \mathcal{X}$ there exists $\widetilde{P}_{C,\vec{Y}}(\cdot \mid w,x,z) \in \Delta(\mathcal{C} \times \mathcal{Y} \times \mathcal{Y})$ satisfying*

i. For all $c \in \{0,1\}$, $\tilde{c} \neq c$,

$$\sum_{\vec{y}} \widetilde{P}_{C,\vec{Y}}(c,\vec{y} \mid w,x,z) U(c,\vec{y};w,z) \geq \sum_{\vec{y}} \widetilde{P}_{C,\vec{Y}}(c,\vec{y} \mid w,x,z) U(\tilde{c},\vec{y};w,z).$$

ii. For all $y \in \mathcal{Y}$, $\sum_{y_1} \widetilde{P}_{C,\vec{Y}}(0,y,y_1 \mid w,x,z) = P_{C,Y_0}(0,y \mid w,x,z)$ and $\sum_{y_0} \widetilde{P}_{C,\vec{Y}}(1,y_0,y \mid w,x,z) = P_{C,Y_1}(1,y \mid w,x,z)$.

iii. For all $\tilde{z} \in \mathcal{Z}$, $\vec{y} \in \mathcal{Y} \times \mathcal{Y}$,

$$0 \leq \widetilde{P}_{C,\vec{Y}}(1,\vec{y} \mid w,x,z) + \widetilde{P}_{C,\vec{Y}}(0,\vec{y} \mid w,x,z) \leq P_{C,Y_1}(1,y_1 \mid w,x,\tilde{z}) + P_{C,Y_0}(0,y_0 \mid w,x,\tilde{z}).$$

For each $c \in \{0,1\}$ and $\tilde{c} \neq c$, define the $1 \times N_y^2$ dimensional row vector $A^c_{w,x,z}(U)$ as

$$A^c_{w,x,z}(U) = \left(U(\tilde{c},\vec{y}_1;w,z) - U(c,\vec{y}_1;w,z), \ldots, U(\tilde{c},\vec{y}_{N_y};w,z) - U(c,\vec{y}_{N_y};w,z)\right).$$

For each $(w,x) \in \mathcal{W} \times \mathcal{X}$, define the $2 \times 2N_y^2$ dimensional block diagonal matrix $A_{w,x,z}(U)$ as

$$A_{w,x,z}(U) = \begin{pmatrix} A^0_{w,x,z}(U) & \\ & A^1_{w,x,z}(U) \end{pmatrix}.$$

Define the $2d_w d_x \times 2d_w d_x N_y^2$ dimensional block diagnoal matrix $A_z(U)$ as

$$A_z(U) = \begin{pmatrix} A_{w_1,x_1,z}(U) & & & \\ & A_{w_1,x_2,z}(U) & & \\ & & \ddots & \\ & & & A_{w_{d_w},x_{d_x},z}(U) \end{pmatrix}.$$

Letting $\widetilde{P}_{C,\vec{Y}}(\cdot \mid z) = \left(P_{C,\vec{Y}}(\cdot \mid w_1,x_1,z), \ldots, P_{C,\vec{Y}}(\cdot \mid w_{d_w},x_{d_x},z)\right)$, the revealed preference constraints in (i) of Lemma D.1 can be rewritten as

$$A_z(U)\widetilde{P}_{C,\vec{Y}}(\cdot \mid z) \leq 0,$$

where $\widetilde{P}_{C,\vec{Y}}(\cdot \mid z)$ is a $2d_w d_x N_y^2 \times 1$ dimensional vector.

We may construct $2d_w d_x N_y \times 2d_w d_x N_y^2$ dimensional matrix $B_{z,eq}$ that forms the data consistency conditions in (ii) of Lemma D.1. For each $\tilde{z} \in \mathcal{Z}$, we may construct the $d_w d_x N_y^2 \times 2d_w d_x N_y^2$ matrices $\overline{B}_{z,\tilde{z}}$ that forms the upper bounds in (iii) of Lemma D.1. Stack these together to form $\overline{B}_z$, Finally, define $d_w \cdot \times 2d_w d_x N_y^2$ matrix $D_{z,eq}$ that imposes $\widetilde{P}_{C,\vec{Y}}(\cdot \mid w,x,z)$ sums to one and the $2d_w d_x N_y^2 \times 2d_w d_x N_y^2$ matrix $D_{z,+}$ that imposes each element of $\widetilde{P}_{C,\vec{Y}}(\cdot \mid w,x,z)$ is non-negative.

We next introduce non-negative slack parameters associated with each inequality constrain in

(ii)-(iii), and rewrite (ii)-(iii) in Lemma D.1 as

$$\underbrace{\begin{pmatrix} B_{z,eq} & 0 \\ \overline{B}_z & 1 \end{pmatrix}}_{:=B_z} \underbrace{\begin{pmatrix} \widetilde{P}_{C,\vec{Y}}(\,\cdot \mid z) \\ \overline{s}_z \end{pmatrix}}_{:=\delta} = \mu(P)$$

where $\mu(P)$ is the vector that collects together the observable data and the bounds, and δ is a $2d_w d_x N_y^2 + d_w d_x N_y^2(N_z - 1)$ dimensional vector.

Therefore, we observe that Lemma D.1 implies that the decision maker's choices at $z \in \mathcal{Z}$ are consistent with expected utility maximization behavior at utility function U if and only if there exists vector δ satisfying

$$\underbrace{\begin{pmatrix} A_z(U) & 0 \\ D_{z,eq} & 0 \\ -D_{z,eq} & 0 \\ D_{z,+} & 0 \\ 0 & D_{z,+} \end{pmatrix}}_{:=\tilde{A}_z(U)} \delta \leq \underbrace{\begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}}_{:=b} \text{ and } B_z\delta = \mu(P).$$

The matrix B_z has full row rank, $\text{nrow}(B_z) = 2d_w d_x N_y + d_w d_x N_y^2(N_z - 1)$, $\text{ncol}(B_z) = 2d_w d_x N_y^2 + d_w d_x N_y^2(N_z - 1)$, and so $\text{nrow}(B_z) \leq \text{ncol}(B_z)$. Therefore, we may define $H_z = \begin{pmatrix} B_z \\ \Gamma_z \end{pmatrix}$ to be the full-rank, square matrix that pads B_z with linearly independent rows. Then,

$$\tilde{A}_z(U)\delta = \tilde{A}_z(U)H_z^{-1}H_z\delta = \tilde{A}_z(U)H_z^{-1}\begin{pmatrix} \mu(P) \\ \tilde{\delta} \end{pmatrix},$$

where $\tilde{\delta} := \Gamma_z\delta$ is a $2d_w d_x N_y(N_y - 1)$ dimensional vector. This completes the proof with the slight abuse of notation by also defining $\tilde{A}_z(U)H_z^{-1}$ to be $\tilde{A}(U)$. □

Proposition D.2 showed that testing whether the decision maker's choices are consistent with expected utility maximization behavior at a candidate utility function may be reduced to testing a many moment inequalities with linear nuisance parameters. This same testing problem may be also reduced to testing whether there exists a non-negative solution to a large, linear system of equations, which was recently studied in Kitamura and Stoye (2018) and Fang et al. (2020).

For each $w \in \mathcal{W}$, define $D_w \colon \mathcal{X} \to \{1, \ldots, N_d\}$ to be some function that partitions the support of the characteristics $x \in \mathcal{X}$ into the level sets $\{x \colon D_w(x) = d\}$. Through an application of iterated expectations, if the decision maker's choices are consistent with expected utility maximization behavior at some utility function U, then their choices must satisfy *implied* revealed preference inequalities.

Corollary D.1. *Suppose the decision maker's choices are consistent with expected utility maximization behavior at some utility function U. Then, for all $w \in \mathcal{W}$ and $d \in \{1, \ldots, N_d\}$,*

$$\mathbb{E}_Q\left[U(c, \vec{Y}; W) \mid C = c, W = w, D_w(X) = d\right] \geq \mathbb{E}_Q\left[U(c', \vec{Y}; W) \mid C = c, W = w, D_w(X) = d\right],$$

for all $c \in \{0,1\}$, $(w,d) \in \mathcal{W} \times \{1,\ldots,N_d\}$ *with* $\pi_c(w,d) := P(C=c \mid W=w, D_w(X)=d) > 0$ *and* $c' \neq c$, *where* $\mathbb{E}_Q[\cdot]$ *is the expectation under* Q *and for some* $\widetilde{P}_{\vec{Y}}(\cdot \mid 0,w,x) \in \mathcal{B}_{0,w,x}$, $\widetilde{P}_{Y^*}(\cdot \mid 1,w,x) \in \mathcal{B}_{1,w,x}$ *such that*

$$Q(W=w, D_w(X)=d, C=c, \vec{Y}=\vec{y}) = \sum_{x:\, D_w(x)=d} \widetilde{P}_{\vec{Y}}(\vec{y} \mid c,w,x) P(c,w,x).$$

Therefore, in treatment decisions, researchers may test a lower dimensional set of moment inequalities with linear nuisance parameters that characterize the set of implied revealed preference inequalities. This is useful as recent work develops computationally tractable and power inference procedures for lower-dimensional moment inequality with linear nuisance parameters such as Andrews, Roth and Pakes (2019), Cox and Shi (2020) and Rambachan and Roth (2020).

D.3 Expected Social Welfare Under the Decision Maker's Observed Choices

Consider a policymaker with social welfare function $U^*(0,0) < 0, U^*(1,1) < 0$ as in Section 6.1. Total expected social welfare under the decision maker's observed choices is given by

$$\begin{aligned}\theta^{DM}(U^*) =& U^*(1,1)P(C=1,Y^*=1) + U^*(0,0)P(C=0) \\ &- U^*(0,0) \sum_{(w,x)\in\mathcal{W}\times\mathcal{X}} P_{C,Y^*}(0,1 \mid w,x) P(w,x).\end{aligned}$$

Since $P_{C,Y^*}(0,1 \mid w,x)$ is partially identified, total expected social welfare under the decision maker's observed choices is also partially identified and the sharp identified set is an interval.

Proposition D.3. *Consider a screening decision with a binary outcome and a policymaker with social welfare function* $U^*(0,0) < 0, U^*(1,1) < 0$. *The sharp identified set of total expected social welfare under the decision maker's observed choices, denoted by* $\mathcal{H}_P(\theta^{DM}(U^*); \mathcal{B})$, *is an interval with* $\mathcal{H}_P(\theta^{DM}(U^*); \mathcal{B}) = \left[\underline{\theta}^{DM}(U^*), \overline{\theta}^{DM}(U^*)\right]$, *where*

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

where

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

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

The sharp identified set of the conditional probability of $C=0, Y^*=1$ given the characteristics is an interval, and therefore the sharp identified set of total expected social welfare under the decision maker's observed choices can be characterized as the solution to two linear programs. Provided the joint distribution of the characteristics (W,X) are known, then testing the null hypothesis that total expected social welfare is equal to some candidate value is equivalent to testing a system of

moment inequalities with a large number of nuisance parameters that enter the moments linearly. A confidence interval for total expected social welfare under the decision maker's observed choices can be constructed through test inversion.

Proposition D.4. *Consider a binary screening decision and a policymaker with social welfare function $U^*(0,0) < 0, U^*(1,1) < 0$. Conditional on the characteristics (W, X), testing the null hypothesis $H_0\colon \theta^{DM}(U^*) = \theta_0$ is equivalent to testing whether*

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

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

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

$$-U^*(0,0) \sum_{(w,x)\in\mathcal{W}\times\mathcal{X}} \widetilde{P}_{C,Y^*}(0,1 \mid w,x) P(W=w,X=x) =$$

$$\theta_0 - U^*(1,1)P(C=1,Y^*=1) - U^*(0,0)P(C=0)$$

and for each $(w,x) \in \mathcal{W}\times\mathcal{X}$

$$\underline{P}(C=0,Y^*=1 \mid W=w,X=x) \leq \widetilde{P}_{C,Y^*}(0,1 \mid w,x) \leq \overline{P}(C=0,Y^*=1 \mid W=w,X=x).$$

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

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

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

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

where $\Gamma\widetilde{P}_{C,Y^*}(0,1) = \begin{pmatrix} \Gamma_{(1,\cdot)}\widetilde{P}_{C,Y^*}(0,1) \\ \Gamma_{(-1,\cdot)}\widetilde{P}_{C,Y^*}(0,1) \end{pmatrix} = \begin{pmatrix} \theta_0 - U^*(1,1)P(C=1,Y^*=1) - U^*(0,0)P(C=0) \\ \delta \end{pmatrix}$

defining $\delta = \Gamma_{(-1,\cdot)}\widetilde{P}_{C,Y^*}(0,1)$ and $\widetilde{A} = A\Gamma^{-1}$. $\square$

E Proofs of Main Results

Proof of Theorem 2.1

I prove the following Lemma, and then show that it implies Theorem 2.1.

Lemma E.1. *The decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a utility function $U \in \mathcal{U}$, $\widetilde{P}_{\vec{Y}}(\cdot \mid 0,w,x) \in \mathcal{B}_{0,w,x}$ and $\widetilde{P}_{\vec{Y}}(\cdot \mid 1,w,x) \in \mathcal{B}_{1,w,x}$ that satisfies for all $(w,x) \in \mathcal{W} \times \mathcal{X}$, $c \in \{0,1\}$, $c' \neq c$,*

$$\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} \widetilde{P}_{\vec{Y}}(\vec{y} \mid c,w,x) P_C(c \mid w,x) U(c,\vec{y};w) \geq \sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} \widetilde{P}_{\vec{Y}}(\vec{y} \mid c,w,x) P_C(c \mid w,x) U(c',\vec{y};w,).$$

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

First, I show that if the decision maker's choices are consistent with expected utility maximization behavior at some utility function U, joint distribution $(W,X,V,C,\vec{Y}) \sim Q$ and private information with support $\mathcal{V}$, then her choices are also consistent with expected utility maximization behavior at some finite support private information. Partition the original signal space $\mathcal{V}$ into the subsets $\mathcal{V}_{\{0\}}, \mathcal{V}_{\{1\}}, \mathcal{V}_{\{0,1\}}$, which collect together the signals $v \in \mathcal{V}$ at which the decision maker strictly prefers $C=0$, strictly prefers $C=1$ and is indifferent between $C=0, C=1$ respectively. Define the finite support signal space $\widetilde{\mathcal{V}} = \{v_{\{0\}}, v_{\{1\}}, v_{\{0,1\}}\}$ and the finite support private information $\widetilde{V} \in \widetilde{\mathcal{V}}$ as

$$\begin{aligned}
\widetilde{Q}(\widetilde{V} = v_{\{0\}} \mid \vec{Y} = \vec{y}, W = w, X = x) &= Q(V \in \mathcal{V}_{\{0\}} \mid \vec{Y} = \vec{y}, W = w, X = x) \\
\widetilde{Q}(\widetilde{V} = v_{\{1\}} \mid \vec{Y} = \vec{y}, W = w, X = x) &= Q(V \in \mathcal{V}_{\{1\}} \mid \vec{Y} = \vec{y}, W = w, X = x) \\
\widetilde{Q}(\widetilde{V} = v_{\{0,1\}} \mid \vec{Y} = \vec{y}, W = w, X = x) &= Q(V \in \mathcal{V}_{\{0,1\}} \mid \vec{Y} = \vec{y}, W = w, X = x).
\end{aligned}$$

Define $\widetilde{Q}(C=0 \mid \widetilde{V} = v_{\{0\}}, W = w, X = x) = 1$, $\widetilde{Q}(C=1 \mid \widetilde{V} = v_{\{1\}}, W = w, X = x) = 1$ and

$$\widetilde{Q}(C=1 \mid \widetilde{V} = v_{\{0,1\}}, W = w, X = x) = \frac{Q(C=1, V \in \mathcal{V}_{\{0,1\}} \mid W = w, X = x)}{Q(V \in \mathcal{V}_{\{0,1\}} \mid W = w, X = x)}.$$

Define the finite support expected utility representation for the decision maker by the utility function U and the random vector $(W,X,\widetilde{V},C,\vec{Y}) \sim \widetilde{Q}$, where $\widetilde{Q}(w,x,v,c,\vec{y}) = Q(w,x,\vec{y})\widetilde{Q}(\widetilde{v} \mid w,x,\vec{y})\widetilde{Q}(c \mid w,x,\widetilde{v})$. The information set and expected utility maximization conditions are satisfied by construction. Data consistency is satisfied since it is satisfied at the original private infor-

mation $V \in \mathcal{V}$. To see this, notice that for all $(w, x, \vec{y}) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$

$$
\begin{aligned}
&P(C = 1, \vec{Y} = \vec{y} \mid W = w, X = x) = \\
&Q(C = 1, V = \mathcal{V}, \vec{Y} = \vec{y} \mid W = w, X = x) = \\
&Q(C = 1, V \in \mathcal{V}_{\{1\}}, \vec{Y} = \vec{y} \mid W = w, X = x) + Q(C = 1, V \in \mathcal{V}_{\{0,1\}}, \vec{Y} = \vec{y} \mid W = w, X = x) = \\
&\widetilde{Q}(C = 1, \widetilde{V} = v_{C=1}, \vec{Y} = \vec{y} \mid W = w, X = x) + \widetilde{Q}(C = 1, \widetilde{V} = v_{C=e}, \vec{Y} = \vec{y} \mid W = w, X = x) = \\
&\sum_{\widetilde{v} \in \widetilde{\mathcal{V}}} \widetilde{Q}(C = 1, \widetilde{V} = \widetilde{v}, \vec{Y} = \vec{y} \mid W = w, X = x) = \widetilde{Q}(C = 1, \vec{Y} = \vec{y} \mid W = w, X = x).
\end{aligned}
$$

The same argument applies to $P(C = 0, \vec{Y} = \vec{y} \mid W = w, X = x)$. Therefore, for the remainder of the necessity proof, it is without loss of generality to assume the 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 E.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 behavior at some utility function U and joint distribution $(W, X, V, C, \vec{Y}) \sim Q$. Then, for each $c \in \{0, 1\}$, $(w, x, v) \in \mathcal{W} \times \mathcal{X} \times \mathcal{V}$

$$
Q_C(c \mid w, x, v) \left(\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{\vec{Y}}(\vec{y} \mid w, x, v) U(c, \vec{y}; w) \right) \geq Q_C(c \mid w, x, v) \left(\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{\vec{Y}}(\vec{y} \mid w, x, v) U(c', \vec{y}; w) \right)
$$

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

$$
\begin{aligned}
&Q_C(c \mid w, x, v) Q_V(v \mid w, x) \left(\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{\vec{Y}}(\vec{y} \mid w, x, v) U(c, \vec{y}; w) \right) \geq \\
&Q_C(c \mid w, x, v) Q_V(v \mid w, x) \left(\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{\vec{Y}}(\vec{y} \mid w, x, v) U(c', \vec{y}; w) \right).
\end{aligned}
$$

Next, use information set to write $Q_{C,\vec{Y}}(c, \vec{y} \mid w, x, v) = Q_{\vec{Y}}(\vec{y} \mid w, x, v) Q_C(c \mid w, x, v)$ and arrive at

$$
\begin{aligned}
&Q_V(v \mid w, x) \left(\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{C,\vec{Y}}(c, \vec{y} \mid w, x, v) U(c, \vec{y}; w, x) \right) \geq \\
&Q_V(v \mid w, x) \left(\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{C,\vec{Y}}(c, \vec{y} \mid w, x, v) U(c', \vec{y}; w) \right).
\end{aligned}
$$

Finally, we use $Q_{C,\vec{Y}}(c, \vec{y}, v \mid w, x) = Q_{C,\vec{Y}}(c, \vec{y} \mid w, x, v) Q_V(v \mid w, x)$ and then further sum over

$v \in \mathcal{V}$ to arrive at

$$\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} \left(\sum_{v \in \mathcal{V}} Q_{V,C,\vec{Y}}(v, c, \vec{y} \mid w, x) \right) U(c, \vec{y}; w) \geq \sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} \left(\sum_{v \in \mathcal{V}} Q_{V,C,\vec{Y}}(v, c, \vec{y} \mid w, x) \right) U(c', \vec{y}; w)$$

$$\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{C,\vec{Y}}(c, \vec{y} \mid w, x) U(c, \vec{y}; w) \geq \sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{C,\vec{Y}}(c, \vec{y} \mid w, x) U(c', \vec{y}; w).$$

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

Proof of Lemma E.1: Sufficiency To establish sufficiency, I show that if the conditions in Lemma E.1 holds, then private information $v \in \mathcal{V}$ can be constructed that recommends choices $c \in \{0, 1\}$ 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 E.1 are satisfied at some $\widetilde{P}_{\vec{Y}}(\cdot \mid c, w, x) \in \mathcal{B}_{c,w,x}$ for all $c \in \{0, 1\}$, $(w, x) \in \mathcal{W} \times \mathcal{X}$. As notation, let $v \in \mathcal{V} := \{1, \ldots, 3\}$ index the subsets in $2^{\{0,1\}} = \{\{0\}, \{1\}, \{0, 1\}\}$.

For each $(w, x) \in \mathcal{W} \times \mathcal{X}$, define $\mathcal{C}_{w,x} := \{c \colon \pi_c(w, x) > 0\} \subseteq \mathcal{C}$ to be the set of choices selected with positive probability, and partition $\mathcal{C}_{w,x}$ into subsets that have identical choice-dependent potential outcome probabilities. There are $\bar{V}_{w,x} \leq |\mathcal{C}_{w,x}|$ such subsets. Each subset of this partition of $\mathcal{C}_{w,x}$ is a subset in the power set $2^{\{0,1\}}$, and so I associate each subset in this partition with its associated index $v \in \mathcal{V}$. Denote these associated indices by the set $\mathcal{V}_{w,x}$. Denote the choice-dependent potential outcome probability associated with the subset labelled v by $P_{\vec{Y}}(\cdot \mid v, w, x) \in \Delta(\mathcal{Y} \times \mathcal{Y})$. Finally, define $Q_{\vec{Y}}(\vec{y} \mid w, x) = \sum_{c \in \{0,1\}} \widetilde{P}_{\vec{Y}}(\vec{y} \mid c, w, x) \pi_c(w, x)$.

Define the random variable $V \in \mathcal{V}$ according to

$$Q_V(v \mid w, x) = \sum_{c \colon P_{\vec{Y}}(\cdot \mid c, w, x) = P_{\vec{Y}}(\cdot \mid v, w, x)} \pi_c(w, x) \text{ if } v \in \mathcal{V}_{w,x},$$

$$Q_V(v \mid \vec{y}, w, x) = \begin{cases} \frac{\widetilde{P}_{\vec{Y}}(\vec{y} \mid v, w, x) Q(V = v \mid w, x)}{Q_{\vec{Y}}(\vec{y} \mid w, x)} \text{ if } v \in \mathcal{V}_{w,x} \text{ and } Q_{\vec{Y}}(\vec{y} \mid w, x) > 0, \\ 0 \text{ otherwise.} \end{cases}$$

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

$$Q_C(c \mid v, w, x) = \begin{cases} \pi_c(w, x) \Big/ \left(\sum_{\tilde{c} \colon P_{\vec{Y}}(\cdot \mid \tilde{c}, w, x) = P_{\vec{Y}}(\cdot \mid v, w, x)} \pi_{\tilde{c}}(w, x) \right) \text{ if } v \in \mathcal{V}_{w,x} \text{ and } P_{\vec{Y}}(\cdot \mid c, w, x) = \\ \qquad P_{\vec{Y}}(\cdot \mid v, w, x) \\ 0 \text{ otherwise.} \end{cases}$$

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

$$Q(w, x, \vec{y}, v, c) = P_{W,X}(w, x) Q_{\vec{Y}}(\vec{y} \mid w, x) Q_V(V = v \mid \vec{y}, w, x) Q_C(c \mid v, w, x).$$

We now check that this construction satisfies information set, expected utility maximization and data consistency. First, information set is satisfied since $Q_{C,\vec{Y}}(c, \vec{y} \mid w, x, v) = Q_{\vec{Y}}(\vec{y} \mid$

$w, x, v)Q_C(c \mid w, x, v)$ by construction. Next, for any $(w, x) \in \mathcal{W} \times \mathcal{X}$ and $c \in \mathcal{C}_{w,x}$, define $v_{c,w,x} \in \mathcal{V}_{w,x}$ to be the label satisfying $P_{\vec{Y}}(\cdot \mid c, w, x) = P_{\vec{Y}}(\cdot \mid v, w, x)$. For $P_{C,\vec{Y}}(c, \vec{y} \mid w, x) > 0$, observe that

$$
\begin{aligned}
&P_{C,\vec{Y}}(c, \vec{y} \mid w, x) = \\
&\widetilde{P}_{\vec{Y}}(\vec{y} \mid c, w, x) P_C(c \mid w, x) = \\
&Q_{\vec{Y}}(\vec{y} \mid w, x) \frac{Q_{\vec{Y}}(\vec{y} \mid v_{c,w,x}, w, x) \sum_{\tilde{c}:\ \frac{P_{\vec{Y}}(\cdot \mid \tilde{c}, w, x)}{P_{\vec{Y}}(\cdot \mid \tilde{c}, w, x)} = } P_C(\tilde{c} \mid w, x)}{Q_{\vec{Y}}(\vec{y} \mid w, x)} \frac{P_C(c \mid w, x)}{\sum_{\tilde{c}:\ \frac{P_{\vec{Y}}(\cdot \mid \tilde{c}, w, x)}{P_{\vec{Y}}(\cdot \mid \tilde{c}, w, x)} = } P_C(\tilde{c} \mid w, x)} = \\
&Q_{\vec{Y}}(\vec{y} \mid w, x) Q_V(v_{c,w,x} \mid \vec{y}, w, x) Q_C(c \mid v_{c,w,x}, w, x) = \\
&\sum_{v \in \mathcal{V}} Q_{\vec{Y}}(\vec{y} \mid w, x) Q_V(v \mid \vec{y}, w, x) Q_C(c \mid v, w, x) = \sum_{v \in \mathcal{V}} Q_{V,C,\vec{Y}}(v, c, \vec{y} \mid w, x) = Q_{C,\vec{Y}}(c, \vec{y} \mid w, x).
\end{aligned}
$$

Moreover, whenever $P_{C,\vec{Y}}(c, \vec{y} \mid w, x) = 0$, $Q_{\vec{Y}}(\vec{y} \mid v_{c,w,x}, w, x) Q_C(c \mid v_{c,w,x}, w, x) = 0$. Therefore, data consistency holds. Finally, by construction, for $Q_C(C = c \mid V = v_{c,w,x}, W = w, X = x) > 0$,

$$
\begin{aligned}
&Q(\vec{Y} = \vec{y} \mid V = v_{c,w,x}, W = w, X = x) = \\
&\frac{Q(V = v_{c,w,x} \mid \vec{Y} = \vec{y}, W = w, X = x) Q(\vec{Y} = \vec{y} \mid W = w, X = x)}{Q(V = v_{c,w,x} \mid W = w, X = x)} = \\
&\widetilde{P}(\vec{Y} = \vec{y} \mid C = c, W = w, X = x).
\end{aligned}
$$

Therefore, expected utility maximization is satisfied since the inequalities in Lemma E.1 were assumed to hold and data consistency holds.

Lemma E.1 implies Theorem 2.1: Define the joint distribution Q as $Q(w, x, c, \vec{y}) = \widetilde{P}_{\vec{Y}}(\vec{y} \mid c, w, x) P(c, w, x)$. Then, rewrite conditions (i)-(ii) in Lemma E.1 as: for all $c \in \{0, 1\}$ and $c' \neq c$,

$$
\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{C,\vec{Y}}(c, \vec{y} \mid w, x) U(c, \vec{y}; w) \geq \sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{C,\vec{Y}}(c, \vec{y} \mid w, x) U(c', \vec{y}; w).
$$

Notice that if $P_C(c \mid w, x) = 0$, then $Q_{C,\vec{Y}}(c, \vec{y} \mid w, x) = 0$. Therefore, the inequalities involving $c \in \mathcal{C}$ with $\pi_c(w, x) = 0$ are satisfied. Next, inequalities involving $c \in \mathcal{C}$ with $\pi_c(w, x) > 0$ can be equivalently rewritten as

$$
\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{C,\vec{Y}}(\vec{y} \mid c, w, x) U(c, \vec{y}; w) \geq \sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{C,\vec{Y}}(\vec{y} \mid c, w, x) U(c', \vec{y}; w).
$$

The statement of Theorem 2.1 follows by noticing that

$$\sum_{\vec{y}\in\mathcal{Y}\times\mathcal{Y}} Q_{\vec{Y}}(\vec{y} \mid c,w,x)U(c,\vec{y};w) = \mathbb{E}_Q\left[U(c,\vec{Y};w) \mid C=c, W=w, X=x\right],$$
$$\sum_{\vec{y}\in\mathcal{Y}\times\mathcal{Y}} Q_{Y^*}(\vec{y} \mid c,w,x)U(c',\vec{y};w) = \mathbb{E}_Q\left[U(c',\vec{Y};w) \mid C=c, W=w, X=x\right].$$

□

Proof of Theorem 3.1

Lemma E.2. *The decision maker's choices are consistent with expected utility maximization behavior at some strict preference utility function if and only if there exists strict preference utility functions U satisfying*

i. $P_{Y^*}(1 \mid 1,w,x) \leq \frac{U(0,0;w)}{U(0,0;w)+U(1,1;w)}$ *for all* $(w,x) \in \mathcal{W}\times\mathcal{X}$ *with* $\pi_1(w,x) > 0$,

ii. $\frac{U(0,0;w)}{U(0,0;w)+U(1,1;w)} \leq \overline{P}_{Y^*}(1 \mid 0,w,x)$ *for all* $(w,x) \in \mathcal{W}\times\mathcal{X}$ *with* $\pi_0(w,x) > 0$.

Proof. This is an immediate consequence of applying Lemma E.1 to analyzing expected utility maximization at strict preferences in a screening decision with a binary outcome. For all $(w,x) \in \mathcal{W}\times\mathcal{X}$ with $\pi_1(w,x) > 0$, Lemma E.1 requires $P_{Y^*}(1 \mid 1,w,x) \leq \frac{U(0,0;w)}{U(0,0;w)+U(1,1;w)}$. For all $(w,x) \in \mathcal{W}\times\mathcal{X}$ with $\pi_0(w,x) > 0$, Lemma E.1 requires $\frac{U(0,0;w)}{U(0,0;w)+U(1,1;w)} \leq P_{Y^*}(1 \mid 0,w,x)$. Applying the bounds $\underline{P}_{Y^*}(1 \mid 0,w,x) \leq P_{Y^*}(1 \mid 0,w,x) \leq \overline{P}_{Y^*}(1 \mid 0,w,x)$ then delivers the result. □

By Lemma E.2, the human DM's choices are consistent with expected utility maximization behavior if and only if there exists strict preference utility functions U satisfying

$$\max_{x\in\mathcal{X}^1(w)} P_{Y^*}(1 \mid 1,w,x) \leq \frac{U(0,0;w)}{U(0,0;w)+U(1,1;w)} \leq \min_{x\in\mathcal{X}^0(w)} \overline{P}_{Y^*}(1 \mid 0,w,x)$$

for all $w \in \mathcal{W}$. These inequalities are only non-empty if the stated conditions in Theorem 3.1 are satisfied. The characterization of the identified set of utility functions also follows from Lemma E.2. □

Proof of Proposition 3.1

Under Assumption 3.1, $P_{Y^*}(1 \mid w,x,z) = P_{Y^*}(1 \mid w,x,\tilde{z}) = P_{Y^*}(1 \mid w,x)$ for all $(w,x) \in \mathcal{W}\times\mathcal{X}$ and $z,\tilde{z} \in \mathcal{Z}$. Furthermore, $P_{Y^*}(1 \mid w,x)$ is bounded above and below by

$$P_{C,Y^*}(1,1 \mid w,x,z) \leq P_{Y^*}(1 \mid w,x) \leq \pi_0(w,x,z) + P_{C,Y^*}(1,1 \mid w,x,z)$$

for all $z \in \mathcal{Z}$.

These bounds on $P_{Y^*}(1 \mid w,x)$ are sharp by a simple application of Arstein's Theorem. I drop the conditioning on $(w,x) \in \mathcal{W}\times\mathcal{X}$ to simplify notation, and so this argument applies conditionally on each $(w,x) \in \mathcal{W}\times\mathcal{X}$. The screening decision setting establishes the model correspondence $G\colon \mathcal{Y} \to \mathcal{Z}\times\mathcal{C}\times(\mathcal{Y}\cup\{0\})$, where $G(y^*) = \{(z,c,y)\colon y = y^*1\{c=1\}\}$. The

reverse correspondence is given by $G^{-1}(z,c,y) = \{y^* \colon y = y^* 1\{c=1\}\}$. The observable joint distribution $(Z,C,Y) \sim P$ characterizes a random set $G^{-1}(Z,C,Y)$ via the generalized likelihood $T(A \mid Z=z) = P\left((C,Y) \colon G^{-1}(z,C,Y) \cap A \neq \emptyset\right)$ for all $A \in 2^{\mathcal{Y}}$. Artstein's Theorem implies that there exists a random variable Y^* that rationalizes the observed data through the model correspondence G if and only if there exists some $Y^* \sim \widetilde{P}$ satisfying

$$\widetilde{P}(A) \leq T(A \mid Z=z) \text{ for all } A \in 2^{\mathcal{Y}} \text{ and } z \in \mathcal{Z}.$$

Let $\widetilde{\mathcal{P}}_{Y^*}(\cdot \mid z)$ be the set of distributions on $\mathcal{Y}$ that satisfy these inequalities at a given $z \in \mathcal{Z}$. A sharp characterization of identified set for the marginal distribution of Y^* is then given by $\mathcal{H}_P(P_{Y^*}(\cdot)) = \bigcap_{z \in \mathcal{Z}} \widetilde{\mathcal{P}}_{Y^*}(\cdot \mid z)$. For $\mathcal{Y} = \{0,1\}$. these inequalities give for each $z \in \mathcal{Z}$,

$$\begin{aligned}
\widetilde{P}(Y^*=0) &\leq P(C=0 \mid Z=z) + P(C=1, Y=0 \mid Z=z) \\
\widetilde{P}(Y^*=1) &\leq P(C=0 \mid Z=z) + P(C=1, Y=1 \mid Z=z).
\end{aligned}$$

Since $\widetilde{P}(Y^*=0) + \widetilde{P}(Y^*=1) = 1$, these inequalities may be further rewritten as requiring for each $z \in \mathcal{Z}$

$$P(C=1,Y=1 \mid Z=z) \leq \widetilde{P}(Y^*=1) \leq P(C=0 \mid Z=z) + P(C=1,Y=1 \mid Z=z).$$

This delivers sharp bonds on the marginal distribution of Y^* conditional on any $z \in \mathcal{Z}$ since the instrument is assumed to be independent of the outcome of interest. The sharpness of the bounds on $P(C=0, Y^*=1 \mid W=w, X=x, Z=z)$ immediately follows since $P(C=1, Y^*=1 \mid W=w, X=x, Z=z) = P(C=1, Y=1 \mid W=w, X=x)$ are observed. □

Proof of Theorem 4.1

To prove this result, I first establish the following lemma, and then show Theorem 4.1 follows as a consequence.

Lemma E.3. *Assume $\widetilde{P}_{\vec{Y}}(\cdot \mid w,x) > 0$ for all $\widetilde{P}_{\vec{Y}}(\cdot \mid w,x) \in \mathcal{H}_P(P_{\vec{Y}}(\cdot \mid w,x); \mathcal{B}_{w,x})$ and all $(w,x) \in \mathcal{W} \times \mathcal{X}$. 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 $Q_{\vec{Y}}(\cdot \mid w,x) \in \Delta(\mathcal{Y} \times \mathcal{Y})$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$, $\widetilde{P}_{\vec{Y}}(\cdot \mid 0,w,x) \in \mathcal{B}_{0,w,x}$, $\widetilde{P}_{\vec{Y}}(\cdot \mid 1,w,x) \in \mathcal{B}_{1,w,x}$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$ satisfying for all $c \in \{0,1\}, c' \neq c$*

$$\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{\vec{Y}}(\vec{y} \mid w,x) \widetilde{P}_C(c \mid \vec{y},w,x) U(c,\vec{y};w) \geq \sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} Q_{\vec{Y}}(\vec{y} \mid w,x) \widetilde{P}_C(c \mid \vec{y},w,x) U(c',\vec{y};w),$$

where $\widetilde{P}_C(c \mid \vec{y},w,x) = \frac{\widetilde{P}_{\vec{Y}}(\vec{y} \mid c,w,x)\pi_c(w,x)}{\widetilde{P}_{\vec{Y}}(\vec{y} \mid w,x)}$ *and* $\widetilde{P}_{\vec{Y}}(\vec{y} \mid w,x) = \widetilde{P}_{\vec{Y}}(\vec{y} \mid 0,w,x)\pi_0(w,x) + \widetilde{P}_{\vec{Y}}(\vec{y} \mid 1,w,x)\pi_1(w,x)$.

Proof of Lemma E.3: Necessity To show necessity, we apply the same steps as the proof of necessity for Lemma E.1. First, by an analogous argument as given in the proof of necessity for Lemma E.1, it is without loss of generality to assume the private information $V \in \mathcal{V}$ has finite

support. Second, following the same steps as the proof of necessity for Lemma E.1, I arrive at

$$\sum_{\vec{y}\in\mathcal{Y}\times\mathcal{Y}} Q_{C,\vec{Y}}(c,\vec{y}\mid w,x)U(c,\vec{y};w) \geq \sum_{\vec{y}\in\mathcal{Y}\times\mathcal{Y}} Q_{C,\vec{Y}}(c,\vec{y},w,x)U(c',\vec{y};w).$$

Then, we immediately observe that $Q_{C,\vec{Y}}(c,\vec{y}\mid w,x) = Q_C(c\mid\vec{y},w,x)Q_{\vec{Y}}(\vec{y}\mid w,x) = \widetilde{P}_C(c\mid\vec{y},w,x)Q_{\vec{Y}}(\vec{y}\mid w,x)$, where the last equality follows via data consistency with inaccurate beliefs.

Proof of Lemma E.3: Sufficiency To show sufficiency, suppose that the conditions in Lemma E.3 are satisfied at some $\widetilde{P}_{\vec{Y}}(\cdot\mid c,w,x)\in\mathcal{B}_{c,w,x}$ for $c\in\{0,1\}$, $(w,x)\in\mathcal{W}\times\mathcal{X}$ and some $Q_{\vec{Y}}(\cdot\mid w,x)\in\Delta(\mathcal{Y}\times\mathcal{Y})$ for all $(w,x)\in\mathcal{W}\times\mathcal{X}$.

Define the joint distribution $(W,X,C,\vec{Y})\sim\widetilde{P}$ according to $\widetilde{P}(w,x,c,\vec{y}) = \widetilde{P}_C(c\mid\vec{y},w,x)Q_{\vec{Y}}(\vec{y}\mid w,x)P(w,x)$, where $\widetilde{P}_C(\cdot\mid\vec{y},w,x)$ is defined in the statement of the Lemma. Given the inequalities in the Lemma, we can construct a joint distribution $(W,X,V,C,\vec{Y})\sim Q$ to satisfy information set, expected utility maximization behavior and data consistency (Definition 3) for the constructed joint distribution $(W,X,C,\vec{Y})\sim\widetilde{P}$ following the same sufficiency argument as given in Lemma E.1. This constructed joint distribution $(W,X,V,C,\vec{Y})\sim Q$ will be an expected utility maximization representation under inaccurate beliefs.

As notation, define $\widetilde{P}_C(c\mid w,x)$ to be the probability of $C=c$ given $W=w,X=x$ and $\widetilde{P}_{\vec{Y}}(\vec{y}\mid c,w,x)$ to be the choice-dependent potential outcome probability given $C=c,W=w,X=x$ under the constructed joint distribution $(W,X,C,\vec{Y})\sim\widetilde{P}$. Let $v\in\mathcal{V}:=\{1,\ldots,3\}$ index the possible subsets in the power set $2^{\{0,1\}}$.

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

Define the random variable $V\in\mathcal{V}$ according to

$$Q(V=v\mid w,x) = \sum_{c\colon \widetilde{P}_{\vec{Y}}(\cdot\mid c,w,x)=\widetilde{P}_{\vec{Y}}(\cdot\mid v,w,x)} \widetilde{P}_C(c\mid w,x) \text{ if } v\in\mathcal{V}_{w,x},$$

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

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

$$Q(C=c\mid v,w,x) = \begin{cases} \frac{\widetilde{P}_C(c\mid w,x)}{\sum_{\tilde{c}\colon \widetilde{P}_{\vec{Y}}(\cdot\mid\tilde{c},w,x)=\widetilde{P}_{\vec{Y}}(\cdot\mid v,w,x)}\widetilde{P}_C(\tilde{c}\mid w,x)} \text{ if } v\in\mathcal{V}_{w,x} \text{ and } \widetilde{P}_{\vec{Y}}(\cdot\mid c,w,x)=\widetilde{P}_{\vec{Y}}(\cdot\mid v,w,x) \\ 0 \text{ otherwise.}\end{cases}$$

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

as

$$Q(w,x,\vec{y},v,c) = P(w,x)Q_{\vec{Y}}(\vec{y} \mid w,x)Q_V(v \mid \vec{y},w,x)Q_C(c \mid v,w,x).$$

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

$$\widetilde{P}_{C,\vec{Y}}(c,\vec{y} \mid w,x) =$$
$$\widetilde{P}_{\vec{Y}}(\vec{y} \mid c,w,x)\widetilde{P}_C(c \mid w,x) =$$
$$Q_{\vec{Y}}(\vec{y} \mid w,x)\frac{Q_{\vec{Y}}(\vec{y} \mid v_{c,w,x},w,x)\sum_{\left\{\tilde{c}:\ \widetilde{P}_{\vec{Y}}(\cdot|\tilde{c},w,x)=\widetilde{P}_{\vec{Y}}(\cdot|v,w,x)\right\}}\widetilde{P}_C(\tilde{c} \mid w,x)}{Q_{\vec{Y}}(\vec{y} \mid w,x)}\frac{\widetilde{P}_C(c \mid w,x)}{\sum_{\left\{\tilde{c}:\ \widetilde{P}_{\vec{Y}}(\cdot|\tilde{c},w,x)=\widetilde{P}_{\vec{Y}}(\cdot|v,w,x)\right\}}\widetilde{P}_C(\tilde{c} \mid w,x)} =$$
$$Q_{\vec{Y}}(y^* \mid w,x)Q_V(v_{c,w,x} \mid \vec{y},w,x)Q_C(c \mid v_{c,w,x},w,x) =$$
$$\sum_{v\in\mathcal{V}} Q_{\vec{Y}}(\vec{y} \mid w,x)Q_V(v \mid \vec{y},w,x)Q_C(c \mid v,w,x) = \sum_{v\in\mathcal{V}} Q(v,c,\vec{y} \mid w,x).$$

Moreover, whenever $\widetilde{P}_{C,\vec{Y}}(c,\vec{y} \mid w,x) = 0$, $Q_{\vec{Y}}(\vec{y} \mid v_{c,w,x},w,x)Q_C(c \mid v_{c,w,x},w,x) = 0$. Therefore, data consistency holds (Definition 3) holds for the constructed joint distribution $(W,X,C,\vec{Y}) \sim \widetilde{P}$. Since $\widetilde{P}_{C,\vec{Y}}(c,y^* \mid w,x) = \widetilde{P}_C(c \mid \vec{y},w,x)Q_{\vec{Y}}(\vec{y} \mid w,x)$ by construction, $(W,X,V,C,\vec{Y}) \sim Q$ satisfies data consistency at inaccurate beliefs (Definition 7). Finally, for $Q(C = c \mid V = v_{c,w,x},W = w,X = x) > 0$,

$$Q(\vec{Y} = \vec{y} \mid V = v_{c,w,x},W = w,X = x) =$$
$$\frac{Q(V = v_{c,w,x} \mid \vec{Y} = \vec{y},W = w,X = x)Q(\vec{Y} = \vec{y} \mid W = w,X = x)}{Q(V = v_{c,w,x} \mid W = w,X = x)} =$$
$$\widetilde{P}(\vec{Y} = \vec{y} \mid C = c,W = w,X = x)$$

and $\widetilde{P}(C = c \mid W = w,X = x) > 0$. Therefore, using data consistency and the inequalities in Lemma E.3, we have that

$$\sum_{\vec{y}\in\mathcal{Y}\times\mathcal{Y}} Q_{\vec{Y}}(\vec{y} \mid v,w,x)U(c,\vec{y};w) \geq \sum_{\vec{y}\in\mathcal{Y}\times\mathcal{Y}} Q_{\vec{Y}}(\vec{y} \mid v,w,x)U(c',\vec{y};w),$$

which follows from the fact that $Q_{\vec{Y}}(\vec{y} \mid w,x)\widetilde{P}_C(c \mid \vec{y},w,x) = Q_{C,\vec{Y}}(c,\vec{y} \mid w,x)$ by data consistency and the construction of $\widetilde{P}$, and $Q(\vec{Y} = \vec{y} \mid V = v_{c,w,x},W = w,X = x) = \widetilde{P}(\vec{Y} = \vec{y} \mid C = c,W = w,X = x)$ as just shown. Therefore, expected utility maximization is also satisfied.

Rewrite inequalities in Lemma E.3 in terms of weights: Define $\widetilde{P}$ as in the statement of the Theorem. Rewrite the condition in Lemma E.3 as: for all $c \in \{0,1\}$ and $\tilde{c} \neq c$,

$$\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} \frac{Q_{\vec{Y}}(\vec{y} \mid w,x)}{\widetilde{P}_{\vec{Y}}(\vec{y} \mid w,x)} \widetilde{P}_{C,\vec{Y}}(c,\vec{y} \mid w,x) U(c,\vec{y};w) \geq \sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} \frac{Q_{\vec{Y}}(\vec{y} \mid w,x)}{\widetilde{P}_{\vec{Y}}(\vec{y} \mid w,x)} \widetilde{P}_{C,\vec{Y}}(c,\vec{y} \mid w,x) U(c',\vec{y};w).$$

Notice that if $\pi_c(w,x) = 0$, then $\widetilde{P}_{C,\vec{Y}}(c,\vec{y} \mid w,x) = 0$. Therefore, the inequalities involving $c \in \{0,1\}$ with $\pi_c(w,x) = 0$ are trivially satisfied. Next, inequalities involving $c \in \{0,1\}$ with $\pi_c(w,x) > 0$ can be equivalently rewritten as

$$\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} \frac{Q_{\vec{Y}}(\vec{y} \mid w,x)}{\widetilde{P}_{\vec{Y}}(\vec{y} \mid w,x)} \widetilde{P}_{\vec{Y}}(\vec{y} \mid c,w,x) U(c,\vec{y};w) \geq \sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} \frac{Q_{\vec{Y}}(\vec{y} \mid w,x)}{\widetilde{P}_{\vec{Y}}(\vec{y} \mid w,x)} \widetilde{P}_{\vec{Y}}(\vec{y} \mid c,w,x) U(c',\vec{y};w).$$

The result follows by noticing that $\sum_{\vec{y} \in \mathcal{Y} \times \mathcal{Y}} \widetilde{P}_{\vec{Y}}(\vec{y} \mid c,w,x) \frac{Q_{\vec{Y}}(\vec{y}|w,x)}{\widetilde{P}_{\vec{Y}}(\vec{y}|w,x)} U(c,\vec{y};w) = \mathbb{E}_{\widetilde{P}}\left[\frac{Q_{\vec{Y}}(\vec{y}|w,x)}{\widetilde{P}_{\vec{Y}}(\vec{y}|w,x)} U(c,\vec{y};w)\right]$ and defining the weights as $\omega(\vec{y};w,x) = \frac{Q_{\vec{Y}}(\vec{y}|w,x)}{\widetilde{P}_{\vec{Y}}(\vec{y}|w,x)}$. □

Proof of Theorem 4.2

Under the stated conditions, if the decision maker's choices are consistent with expected utility maximization behavior at inaccurate beliefs and some strict preference utility function, Theorem 4.1 implies that for all $(w,x) \in \mathcal{W} \times \mathcal{X}$

$$\omega(1;w,x) U(1,1;w,x) P_{Y^*}(1 \mid 1,w,x) \geq \omega(0;w,x) U(0,0;w,x) P_{Y^*}(0 \mid 1,w,x),$$

$$\omega(0;w,x) U(0,0;w,x) \widetilde{P}_{Y^*}(0 \mid 0,w,x) \geq \omega(1;w,x) U(1,1;w,x) \widetilde{P}_{Y^*}(1 \mid 0,w,x),$$

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

$$P_{Y^*}(1 \mid 1,w,x) \leq \frac{\omega(0;w,x) U(0,0;w,x)}{\omega(0;w,x) U(0,0;w,x) + \omega(1;w,x) U(1,1;w,x)} \leq \widetilde{P}_{Y^*}(1 \mid 0,w,x).$$

The result then follows by applying the bounds on $\widetilde{P}_{Y^*}(1 \mid 0,w,x)$ in a screening decision with a binary outcome. □

Proof of Proposition 6.2

First, rewrite $\theta(p^*,U^*)$ as

$$\beta(p^*,U^*) + \ell^{\intercal}(p^*,U^*) P_{C,Y^*}(1,1) + \ell^{\intercal}(p^*,U^*) P_{C,Y^*}(0,1),$$

where $\ell^{\intercal}(p^*,U^*)$ is defined in the statement of the proposition and $P_{C,Y^*}(1,1)$, $P_{C,Y^*}(0,1)$ are the $d_w d_x$ vectors whose elements are the moments $P_{C,Y^*}(1,1 \mid w,x) := P(C=1,Y^*=1 \mid W=w,X=x)$, $P_{C,Y^*}(0,1 \mid w,x) := P(C=0,Y^*=1 \mid W=w,X=x)$ respectively. The null hypothesis $H_0 : \theta(p^*,U^*) = \theta_0$ is equivalent to the null hypothesis that there exists $\widetilde{P}(0,1)$

satisfying

$$\ell^\intercal(p^*, U^*)\widetilde{P}_{C,Y^*}(0,1) = \theta(p^*, U^*) - \beta(p^*, U^*) - \ell^\intercal(p^*, U^*)P_{C,Y^*}(1,1)$$

and, for all $(w, x) \in \mathcal{W} \times \mathcal{X}$,

$$\underline{P}(C = 0, Y^* = 1 \mid W = w, X = x) \leq \widetilde{P}_{C,Y^*}(0, 1 \mid w, x) \leq \overline{P}(C = 0, Y^* = 1 \mid W = w, X = x).$$

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

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

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

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

where $\Gamma\widetilde{P}_{C,Y^*}(0,1) = \begin{pmatrix} \Gamma_{(1,\cdot)}\widetilde{P}_{C,Y^*}(0,1) \\ \Gamma_{(-1,\cdot)}\widetilde{P}_{C,Y^*}(0,1) \end{pmatrix} = \begin{pmatrix} \theta_0 - \beta(p^*, U^*) - \ell^\intercal(p^*, U^*)P_{C,Y^*}(1,1) \\ \delta \end{pmatrix}$ defining $\delta = \Gamma_{(-1,\cdot)}\widetilde{P}_{C,Y^*}(0,1)$ and $\widetilde{A} = A\Gamma^{-1}$.