Predicting and Understanding Individual-Level Choice Under Risk*

Keaton Ellis, Shachar Kariv and Erkut Ozbay†

September 30, 2024

Abstract

We compare the predictive performance of economic models of choice under risk to various machine learning (ML) models by presenting nearly 1,000 subjects with a consumer decision problem—the selection of a bundle of contingent commodities from a budget set. We compare models' predictions at the individual level and relate them to the consistency of decisions with revealed preference axioms. Using dual measures of completeness and restrictiveness, we show that Expected Utility Theory (EUT) performs as well as non-EUT and outperforms all ML models, with a wider margin as choices align more with utility maximization.

*The results reported here were previously distributed in a paper titled "What Can the Demand Analyst Learn from Machine Learning?" The current title draws inspiration from the seminal work by Fudenberg and Liang (2019). We thank Annie Liang for detailed comments and suggestions and Yiting Chen, Emel Filiz-Ozbay, Brian Jabarian, Michael Jordan, Daniel Martin, Yusufcan Masatlioglu, Sendhil Mullainathan, Sara Neff, and Anna Vakarova for helpful conversations. The paper has also benefited from suggestions by the participants of D-TEA (Decision: Theory, Experiments, and Applications), RUD (Risk, Uncertainty, and Decision), WEAI (Western Economic Association International), MLESC24 (Machine Learning in Economics Summer Conference), ESIF-AIML (Economics and AI+ML Meeting), and seminars at several universities. Ellis is grateful for the support from the Foundations of Data Science Institute (FODSI), funded by the National Science Foundation TRIPODS program, and for the hospitality of the Simons Institute for the Theory of Computing at the University of California, Berkeley. The opinions, findings, and conclusions expressed in this material are those of the authors.

†Ellis: University of California, Berkeley (khkellis@berkeley.edu); Kariv: University of California, Berkeley (kariv@berkeley.edu); Ozbay: University of Maryland (ozbay@umd.edu).

JEL Classification Numbers: C63, C91, D81.
Keywords: machine learning, revealed preference, risk preferences, expected utility, non-expected utility, completeness, restrictiveness, experiments.

1 Introduction

Expected Utility Theory (EUT) is central to economics, serving both as a normative guide for choice and also as a descriptive model of how individuals choose. At the same time, much of the experimental and empirical evidence of "anomalies" in choice behavior suggests that EUT may not be the right model of choice under risk. Specifically, empirical violations of the independence axiom on which EUT is based provoke intriguing questions about the rationality of individual choice under risk and, In particular, the status of independence as the touchstone of rationality. These criticisms have led to the development of various non-EUT alternatives that relax the independence axiom while adhering to the basic axioms of ordering (completeness and transitivity) and monotonicity with respect to First-Order Stochastic Dominance (FOSD).

The analysis of choice under risk is, therefore, driven by four key questions commonly posed in demand analysis (Varian, 1982, Varian, 1983): (*i*) *Consistency*. Is choice under risk consistent with a model of utility maximization? (*ii*) *Structure*. Is the observed data consistent with a utility function that aligns with certain theories but not others? (*iii*) *Recoverability*. Can the underlying risk preferences be recovered from observed choices? (*iv*) *Extrapolation*. How accurately can we predict choice under risk in different scenarios?

The economic approach thus involves testing whether behavior can be rationalized by some preference ordering (or posit a utility function with some special structure), deriving the associated demand functions, and fits those to data using some econometric technique. There is a wide variety of formats to this economic approach, ranging from nonparametric to semiparametric to parametric methods. The estimated preference parameters can then be used to extrapolate and predict behavior. By now this type of analysis has become quite standard (Deaton and Muellbauer, 1980).

While economic models revolve around constructing parameter estimates of the underlying utility function and using those to forecast behavior, machine learning

(ML) models are built for the purpose of extrapolation by seeking functions that minimize out-of-sample prediction error. As pointed by Mullainathan and Spiess (2017), among others, ML does not produce stable estimates of the underlying preference parameters. As a result, the "revealed" preference ordering may not be the "true," underlying preference ordering. In that case, positive predictions and welfare conclusions based on the "revealed" preferences will be misleading, at least when applied in other settings. ML, therefore, should be used in economics where improved prediction has large applied value.

This paper explores the promise of ML in predicting *individual-level* choice under risk. We emphasize the term individual to highlight that we will investigate behavior at the level of the individual subject. There is no general reason to suppose that treating aggregate data as if they had been generated by a single type (or a mixture of types) is valid. Clearly, even high-level consistency in individual-level decisions does not imply that aggregate data are consistent. In fact, the considerable heterogeneity in subjects' behaviors entails that even if behaviors are individually consistent, they are mutually inconsistent. Thus, any aggregate-level economic analysis is inevitably misspecified because there is no utility function that pooled choices maximize (Afriat's Theorem).

Most importantly, we present subjects with a standard economic decision problem that can be interpreted either as a portfolio choice problem—the allocation of wealth between two risky assets—or a consumer decision problem—the selection of a bundle of contingent commodities from a standard budget set. These decision problems are presented using a graphical experimental interface of Choi et al. (2007b) that allows for the collection of a rich individual-level data set. Because of the user-friendly interface, each subject faces a large menu of highly heterogeneous budget sets, and the large amount of data generated by this design allows us to apply statistical models to *individual* data rather than pooling data or assuming homogeneity across subjects.[1]

Let $\mathbf{p}^i$ denote the i-th observation of the price vector and $\mathbf{x}^i$ denote the associated demand bundle. Assume we have $i = 1, ..., n$ observations of these prices and quantities generated by some individual's choices. The question we ask (and answer) is which approach—economics or ML—provides the "best estimate" of the

[1]The power of the experiment depends on two factors. The first is that the number of decisions made by each subject is large. The second is that the range of choice sets is generated so that budget lines cross frequently (see Choi et al., 2007b).

demand bundle $\mathbf{x}^0$ when the prevailing prices are $\mathbf{p}^0$ based on previously observed behavior $\{(\mathbf{p}^i, \mathbf{x}^i)\}_{i=1}^n$? The key dual concepts in this regard are *completeness* and *restrictiveness* by Fudenberg et al. (2022, 2023):

- The completeness of a model is the fraction of the predictable variation in the data that the model captures. A more complete model better captures the regularities in the data, but the model might have enough flexibility to accommodate any regularity.

- The restrictiveness of a model discern completeness due to the "right" regularities by evaluating its distance to synthetic data. An unrestrictive model is complete on any possible data, so the fact that it is complete on the actual data is uninstructive.

In the experiment, there are two equiprobable states of nature denoted by $s = 1, 2$ and two associated Arrow securities, each of which promises a token (the experimental currency) payoff in one state and nothing in the other. Let $\mathbf{x} = (x_1, x_2) \geq \mathbf{0}$ denote a bundle of securities, where x_s denotes the number of units of security s. A bundle $\mathbf{x}$ must satisfy the budget constraint $\mathbf{p} \cdot \mathbf{x} = m$, where m is the endowment and $\mathbf{p} = (p_1, p_2) \geq \mathbf{0}$ is the vector of security prices and p_s denotes the price of security s. The data set consists of observations on nearly 1,000 subjects. For each subject, we have 50 observations $\{(\mathbf{p}^i, \mathbf{x}^i)\}_{i=1}^{50}$ over a wide range of budget sets.

For each subject, we first assess, using revealed preference tests, how closely individual choice behavior complies with the Generalized Axiom of Revealed Preference (GARP) and with monotonicity with respect to FOSD (Nishimura et al., 2017 and Polisson et al., 2020). We then calculate the completeness and restrictiveness of the EUT model and a non-EUT model generated by a Rank-Dependent Utility (RDU) function (Quiggin, 1982). RDU weakens the independence axiom but maintains ordering and monotonicity with respect to FOSD, making EUT a special case of this theory.[2]

We find that RDU does not outperform EUT—the average completeness of EUT (89.3%) is essentially the same as that of RDU (89.2%), and the restrictiveness of

[2] Machina (1994) concludes that RDU is "the most natural and useful modification of the classical expected utility formula." Starmer (2000) points out that although the number of non-EUT models "is well into double figures," the preferences generated by RDU is the leading contender. See Diecidue and Wakker (2001) for a comprehensive discussion.

EUT (18.6%) is marginally higher than that of RDU (16.6%). At the individual level, there is considerable heterogeneity in the completeness of the EUT and RDU models across subjects, and notable symmetry between the completeness scores of the two models within subjects. However, EUT has higher completeness for 60.6% of our subjects.[3] We therefore compare EUT to the ML models and replicate all our results—which are nearly identical—with RDU instead of EUT in the Online Appendix.

The core of our analysis involves a *subject-by-subject* comparison of the completeness of EUT with the *most* complete model among *eight* ML models, spanning *three* main families—regularized regressions, tree-based methods, and neural networks. Figure 1 below depicts our main result. The horizontal axis presents quartiles over the distribution of subjects' consistency scores with GARP and FOSD. The vertical axis indicates the fraction of subjects for whom EUT is *more* complete than the *most* complete ML model within each class, as well as *more* complete than the *best* ML model overall (the horizontal lines). Over all subjects, the economic model is more complete for 65.4% and this fraction increases monotonically from 54.2% for subjects in the bottom quartile of consistency scores to 73.8% for subjects in the top quartile, who are (almost) perfectly rationalizable as maximizing a (continuous) utility function that is increasing with respect to FOSD. For those who are generally consistent with GARP and FOSD, there is little room for improving the prediction of the economic models.

[Figure 1 here]

We also note that EUT is not less restrictive than most ML models, which are specifically designed for prediction. Its higher individual-level completeness suggests that EUT is better suited to capturing the heterogeneous behaviors of subjects. Much of the experimental and behavioral literature on decisions under risk focuses on identifying violations of EUT. However, EUT is a fundamental component of economics and should not be discarded lightly, even for the sake of parsimony. We interpret our results as a 'victory' for economic models, particularly EUT, which is foundational to much of economics.

[3]Following the literature, our estimates of completeness come from cross-validation methods. This process conducts out-of-sample estimation, where nested models may outperform their associated nesting models. We discuss this more in Section 3.2.1.

The rest of the paper is organized as follows. The next section provides a discussion of the closely related literature and the main references. Section 3 describes the experimental data and introduces the template for our analysis. Section 4 discusses the results and their importance. Section 5 discusses the contributions that the paper offers, provides directions for future research, and contains some concluding remarks.

2 Related Literature

Our paper contributes to the body of work that seeks to use artificial intelligence (AI) and ML techniques to enhance economic models – theoretical and empirical. We do not attempt to provide a full overview of possible applications.[4] Instead, we focus on closely related papers at the intersection of ML and economic theory.

Peysakhovich and Naecker (2017) compare the performances of EUT and prominent non-EUT alternatives to the performance of regularized regression models using experimental data on the willingness to pay for three-outcome lotteries under risk (known probabilities) and ambiguity (unknown probabilities). While the economic models perform as well as the regularized regression models at predicting choices under risk, they "fail to compete" predicting choices under ambiguity. Fudenberg and Liang (2019) formulate the approach on initial play in 3×3 matrix games. They examine problems where ML models correctly predict (aggregate) modal actions and economic models do not, construct a hypothesis explaining the performance gap, and incorporate their hypothesis via modifications to existing economic theories and successfully close the gap.

Subsequent work applies similar methodologies to other areas of microeconomic theory. Clithero et al. (2023) find a performance gap between the Becker et al. (1964) mechanism and ML models when predicting purchase decisions. Fudenberg and Karreskog Rehbinder (2024) find that semi-grim trigger strategies perform well relative to ML models in predicting cooperation rates in repeated games. Other papers, such as Hsieh et al. (2023) and Peterson et al. (2021), conduct the same

[4] As of September 2024, it is clear that the entirety of economics is fundamentally changed by the introduction of AI and ML, from labor economics (e.g. Brynjolfsson et al., 2023) to macroeconomics (e.g. Fernández-Villaverde et al., 2023), econometrics (e.g. Chernozhukov et al., 2018), and experimental economics (e.g. Horton, 2023). New subfields of human-AI interaction are quickly emerging in behavioral and experimental economics (e.g. Charness et al., 2023; Almog et al., 2024), mechanism design (e.g. Brunnermeier et al., 2023), and others. There also exist applications of economic tools to generative AI models (Chen et al., 2023; Kim et al., 2024).

type of predictive exercise between economic models of choice under risk using neural networks as Bernoulli utility functions.

Two recent papers by Ludwig and Mullainathan (2024) and Mullainathan and Rambachan (2024) investigate closely related questions, albeit with different methodological approaches. Instead of evaluating predictive performance, these papers use generative adversarial approaches that generate synthetic observations to maximize an objective under a constraint that the generations be realistic. While we do not explore these alternative methodologies in the current work, their approach represents promising avenues for future research.

Fudenberg et al. (2022) and Fudenberg et al. (2023) respectively develop the measures of completeness and restrictiveness, which we adopt here to evaluate a model's prediction accuracy and flexibility. Fudenberg et al. (2022) calculate the completeness of models predicting certainty equivalents for binary lotteries under risk (as well as predicting initial play in matrix games and human generation of random sequences). They observe that a three-parameter specification generated by Cumulative Prospect Theory (CPT), proposed by Kahneman and Tversky (1979), is a nearly complete model for predicting their aggregate-level data of certainty equivalents. Building on this analysis, Fudenberg et al. (2023) show that CPT achieves much higher completeness then a two-parameter specification generated by Disappointment Aversion, proposed by Gul (1991), but CPT is also substantially less restrictive. Similarly, Fudenberg and Puri (2022a) and Fudenberg and Puri (2022b) evaluate the completeness of multiple EUT and non-EUT specifications with and without simplicity preferences (Puri, 2018).

We also note that part of the literature focuses on comparing out-of-*domain* performance of economic models against ML models, instead of solely out-of-*sample* performance. This scope is subtly different since the goal is to produce a model that is robust to a shift in the inputs or distribution of responses. In this regard, the literature finds positive results about economic model performance in choice under risk, social preference, and stochastic choice settings.

Andrews et al. (2024) propose two measures of relative cross-domain transfer performance of a model: against (i) optimal cross-domain transfer and (ii) the same model estimated within-domain. They generate statistical methods to estimate transfer performance across domains, showing that economic models overall have better transfer performance than ML models when predicting certainty equivalents of bi-

nary choice lotteries, stemming primarily from their ability to better extrapolate to different payoff values in the domain. Fehr et al. (2023) find similar results in social preferences choice from budget sets when new budget lines fall outside of the training domain. Kobayashi and Lucia (2023) focus on transfer performance on binary prediction tasks in two scenarios with prevalent non-EUT behavior: common ratio tasks á la Allais and preference randomization tasks. They find that ML models have better out-of-sample prediction than three-group economic mixture models within a task type, but worse out-of-domain prediction when training on questions of one type and testing on another.

We share the point of view of Peysakhovich and Naecker (2017), that individual heterogeneity requires behavior to be examined at an individual level, but we go further. Most importantly, previous studies evaluate prediction accuracy and flexibility from a small number of individual decisions and relatively constrained choice scenarios. Aside from pure technicalities, our dataset has a number of advantages over earlier datasets: First, the choice of a bundle subject to a budget constraint provides more information about preferences than a typical discrete choice. Second, we present each subject with many choices, yielding a much larger data set. This makes it possible to analyze behavior at the level of the individual subject, both in terms of prediction as well as estimation, without the need to pool data or assume that subjects are homogeneous. Third, because choices are from standard budget sets, we are able to use classical revealed preference analysis to decide if subject behavior is consistent with the essence of all models of economic decision-making – maximizing a well-behaved utility function – and relate the consistency scores to prediction accuracy at the individual level.

3 Framework for Analysis

3.1 Experiment and Data

In our preferred interpretation of the experiment, there are two equiprobable states of nature $s = 1, 2$ and an Arrow security for each state. Let $x_s \geq 0$ denote the demand for the security that pays off in state s and $p_s > 0$ denote the corresponding price. The budget set is then given by $\mathcal{B} = \{\mathbf{x} : \mathbf{p} \cdot \mathbf{x} = m\}$, where $\mathbf{x} = (x_1, x_2)$ is a demand allocation, $\mathbf{p} = (p_1, p_2)$ is a price vector and m is the endowment. We also define

the *token share* of the security that pays off in one state to be the number of tokens payable in that state as a fraction of the sum of tokens payable in both states, and denote $x = x_1/(x_1 + x_2)$ to be the token share for the first state. Let $\{(\mathcal{B}^i, x^i)\}_{i=1}^n$ be the data generated by a subject's choices from linear budget sets, where $\mathcal{B}^i$ denotes the i-th observation of the budget line and x^i denotes the corresponding token share.[5] Also let $\boldsymbol{B}$ denote the set of budget lines.

The experiment consisted of 50 independent decision problems. In each decision problem, subjects were asked to allocate tokens between two accounts, labeled x and y. The x account corresponds to the x-axis and the y account corresponds to the y-axis in a two-dimensional graph. Each choice involved choosing a point on a budget line of possible token allocations. Each decision problem started by having the computer select a budget line randomly from the set of lines that intersect at least one axis at or above the 50 token level and intersect both axes at or below the 100 token level. The budget lines selected for each subject in his decision problems were independent of each other and of the budget lines selected for other subjects in their decision problems.[6]

To choose an allocation, subjects used the mouse or the arrows on the keyboard to move the pointer on the computer screen to the desired allocation. The payoff at each decision round was determined by the number of tokens in the x account and the number of tokens in the y account. At the end of the round, the computer selected one of the accounts, x or y, with equal probability. Each subject received the number of tokens allocated to the account that was chosen. At the end of the experiment, the computer selected one decision round for each participant and the subject was paid the amount he had earned in that round.

Our subject pool consists of 956 subjects. This dataset includes subjects from

[5] More precisely, the data generated by an individual's choices are $\left\{\left(\bar{x}_1^i, \bar{x}_2^i, x_1^i, x_2^i\right)\right\}_{i=1}^{50}$, where $\left(\bar{x}_1^i, \bar{x}_2^i\right)$ are the endpoints of the budget line and $\left(x_1^i, x_2^i\right)$ are the coordinates of the choice made by the subject and $x_1^i/\bar{x}_1^i + x_2^i/\bar{x}_2^i = 1$ is the the budget line in decision round $i = 1, ...50$. Without loss of generality, the income m is normalized to 1.

[6] Notice that subjects in our experiment could not exhibit almost-optimizing behavior if they had any difficulties understanding the decision problem or using the computer program. The fact that choices nearly satisfy GARP implies that subjects had to exhibit stable patterns of choices over the course of the experiment, which suggests that we maintain subjects' engagement—otherwise, one would expect them to lapse into quasi-random behavior and/or to adopt a low-effort heuristic that would generate many violations (since each round started by having the computer select a budget set randomly, any choice mechanism that does not depend solely on the parameters of the budget set will necessarily generate substantial violations of GARP).

the symmetric treatment of Choi et al. (2007a), similar datasets from subject pools gathered by Zame et al. (2024) and Cappelen et al. (2023), as well as other previously collected data.[7]

3.2 Measures

Following the terminology and notation of Fudenberg et al. (2023), a *predictive mapping* $f : \boldsymbol{B} \to [0,1]$ is a map from budget lines into token shares. Mappings are evaluated using the squared error *loss function* $\ell : [0,1] \times [0,1] \to \mathbb{R}$ where $\ell\left[f(\mathcal{B}^i), x^i\right] = \left[f(\mathcal{B}^i) - x^i\right]^2$ is the error assigned to a predicted token share $f(\mathcal{B}^i)$ when the chosen token share is x^i. The expected prediction error for a mapping f is the expected loss

$$\mathcal{E}_P(f) = \mathbb{E}_P[\ell(f(\mathcal{B}), x)]$$

where P denotes the joint distribution of $(\mathcal{B}, x)$.[8] We are interested in comparing families of parametric mappings $\mathcal{F}_\Theta = \{f_\theta\}_{\theta \in \Theta}$, where the prediction error of a family of parametric mappings $\mathcal{F}_\Theta$ is denoted by the lowest expected prediction error of mappings in the family

$$\mathcal{E}_P(\mathcal{F}_\Theta) = \mathbb{E}_P[\ell(f^*_\Theta(\mathcal{B}^i), x^i)]$$

where $f^*_\Theta = \arg\min_{f \in \mathcal{F}_\Theta} \mathcal{E}_P(f)$.

In recent work, Fudenberg et al. (2022) and Fudenberg et al. (2023) propose a method to use ML techniques to evaluate a theory's prediction accuracy and flexibility. The key dual measures in this regard are *completeness* and *restrictiveness*. The completeness of a model is the fraction of the predictable variation in the data that the model captures. A more complete model better captures the regularities in the data, but the model might have enough flexibility to accommodate any regularity. A more flexible model need not have higher completeness, but such a model is necessarily less parsimonious and thus less falsifiable with an available set of data. The restrictiveness of a model discern completeness due to the "right" regularities by

[7] It is of course possible that presenting choice problems graphically biases behavior in some particular way—every experimental data set may be contaminated by the "frame" that subjects put around the experiment—but there is no evidence that this experimental environment and design are especially vulnerable (see Brown and Healy, 2018 and Azrieli et al., 2018 for these two issues, respectively). The considerable heterogeneity in subjects' behaviors mitigates the framing concerns.

[8] Note that the marginal over budget sets is exogenously set by the experiment, and thus the main object of interest is the (distribution over the) set of responses conditional on a budget set $\mathcal{B}$.

evaluating its distance to synthetic data. An unrestrictive model can be complete on any possible data, so the fact that it is complete on the actual data is uninstructive. The completeness and restrictiveness of *nested* models can be easily compared – the completeness/restrictiveness of a nested model is lower/higher than of the associated nesting model. Yet, in practice, the use of out-of-sample prediction estimates for completeness may result in nested models having a higher completeness.

3.2.1 Completeness

Completeness is the amount that a mapping improves predictions over a *naive* baseline relative to the amount that an *ideal* mapping with *irreducible error* improves predictions over a naive baseline. That is, the completeness of a family of mappings $\mathcal{F}_\Theta$, denoted by κ_Θ, is defined by

$$\kappa_\Theta = \frac{\mathcal{E}_P(f_n) - \mathcal{E}_P(f^*_\Theta)}{\mathcal{E}_P(f_n) - \mathcal{E}_P(f^*)}$$

where f_n is a naive benchmark mapping and the (perfect) predictor with irreducible error is defined by

$$f^*(\mathcal{B}) = \arg\min_{\hat{x}\in[0,1]} \mathbb{E}_P[\ell(\hat{x}, x)|\mathcal{B}].$$

Since subjects see budget sets at most once, it is possible to construct a function from budget sets to demand that will achieve zero error, and thus we assume that $f^*(\mathcal{B}^i) = x^i$. The naive baseline f_n is assumed to be i.i.d uniform choice over the interval $[0, 1]$. Given a subject's true demand x, the expected error of a naive model is $\frac{1}{3}(1 - 3x + 3x^2)$.

We follow Fudenberg et al. (2022) and use 10-fold cross-validation as an estimate of model expected error. In this exercise, the set of individual data $\{(\mathcal{B}^i, x^i)\}_{i=1}^{50}$ is partitioned into ten equally sized, mutually exclusive subsets $Z_1, \ldots, Z_{10}$. Each partition Z_k is then used for out-of-sample prediction, where the complement of the partition Z_{-k} is used to estimate f^*_Θ as $\hat{f}^{-k} = \arg\min_{f_\theta\in\mathcal{F}_\Theta} \frac{1}{45}\sum_{i\notin Z_k} \ell(f_\theta(\mathcal{B}^i), x^i)$. The estimate $\hat{f}^{-k}$ is then used to generate an estimated out-of-sample prediction error over Z_k, $\hat{e}_k = \frac{1}{5}\sum_{i\in Z_k} \ell(\hat{f}^{-k}(\mathcal{B}^i), x^i)$. The estimate of $\mathcal{E}_P(f^*_\Theta)$, denoted $\hat{\mathcal{E}}_\Theta$, is the average of the partition-level error estimates:

$$\hat{\mathcal{E}}_\Theta = \frac{1}{10}\sum_{k=1}^{10} \hat{e}_k$$

The estimate of completeness is thus

$$\hat{\kappa}_\Theta = \frac{\hat{\mathcal{E}}_n - \hat{\mathcal{E}}_\Theta}{\hat{\mathcal{E}}_n - \mathcal{E}_P(f^*)} = \frac{\hat{\mathcal{E}}_n - \hat{\mathcal{E}}_\Theta}{\hat{\mathcal{E}}_n}$$

Fudenberg et al. (2022) show that each individual estimate $\hat{\mathcal{E}}$ is consistent, and thus $\hat{\kappa}_\Theta$ is also consistent. Fudenberg et al. (2023) further extend this - assuming that $\hat{\mathcal{E}}_n > 0$ and regularity conditions, the asymptotic difference between $\hat{\kappa}_\Theta$ and κ_Θ is normal.

3.2.2 Restrictiveness

Restrictiveness is a model-level distance concept which measures the model's flexibility by evaluating the distance of the model to synthetic data. For high completeness models, restrictiveness distinguishes between flexible models that can conform to most mappings f and between models that accurately describe subject behavior. Analyzed together, desirable models are more complete at the individual level and more restrictive at the model level – they explain individual behaviors well, and explain only those behaviors. Let $\mathcal{F}_\mathcal{M}$ denote "permissible mappings" – mappings that are *ex ante* feasible for a decision-maker to have – and let $\mu_{\mathcal{F}_\mathcal{M}}$ denote the uniform distribution over mappings from $\mathcal{F}_\mathcal{M}$. For any two mappings f and f', define the distance between the two functions as

$$d(f, f') = \mathbb{E}_{P_{\boldsymbol{B}}}[\ell(f(\mathcal{B}^i), f'(\mathcal{B}^i)]$$

where $P_{\boldsymbol{B}}$ is the marginal distribution over $\boldsymbol{B}$, and similarly

$$d(\mathcal{F}_\Theta, f') = \inf_{f \in \mathcal{F}_\Theta} d(f, f')$$

is the distance between f' and the closest mapping from $\mathcal{F}_\Theta$. Similar to completeness, restrictiveness is normalized using a naive mapping f_n. Hence, the restrictiveness of a family of mappings $\mathcal{F}_\Theta$, denoted by r_Θ, is defined by

$$r_\Theta = \frac{\mathbb{E}_{\mu_{\mathcal{F}_\mathcal{M}}}[d(\mathcal{F}_\Theta, f)]}{\mathbb{E}_{\mu_{\mathcal{F}_\mathcal{M}}}[d(f_n, f)]}.$$

Like completeness, we use the uniformly random naive benchmark. We let the permissible mappings $\mathcal{F}_M$ be the set of aggregated agents, where a response to a budget line corresponds to a response of a real subject. To generate the distribution $\mu_{\mathcal{F}_\mathcal{M}}$, real subject responses from all 956 subjects are pooled together and partitioned by decile of the price ratio between the cheaper and more expensive good. For each observed budget line, a relative token allocation for the cheaper good is drawn uniformly randomly from that line's decile. The selected allocation may either be $x = x_1/(x_1 + x_2)$ or $1 - x$ depending on which good is cheaper. We group the budget lines by subject, resulting in a set of 956 "representative agents" with synthetic data. Each model is evaluated at the agent level, and the resulting within-sample errors are used to calculate restrictiveness.

3.3 Economic Models

The most basic question to ask about choice data is whether it is consistent with individual utility maximization. If budget sets are linear (as in our preliminary experiment), classical revealed preference theory (Afriat, 1967; Varian, 1982, 1983) provides a direct test: choices in a finite collection of budget sets are consistent with maximizing a well-behaved (that is, piecewise linear, continuous, increasing, and concave) utility function if and only if they satisfy GARP. Hence, in order to decide whether our data are consistent with utility-maximizing behavior we only need to check whether our data satisfies GARP.[9]

However, since GARP offers an exact test—either the data satisfy GARP or they do not—and choice data almost always contain at least some violations, we assess how nearly the data complies with GARP by calculating Afriat (1972) Critical Cost Efficiency Index (CCEI), denoted by e^*. This measures the amount by which each budget constraint must be relaxed in order to remove all violations of GARP. The CCEI is bounded between zero and one, $0 \leq e^* \leq 1$. The closer it is to one, the smaller the perturbation of budget sets required to remove all violations and thus the closer the data are to satisfying GARP.

Beyond consistency, choices can be consistent with GARP and yet fail to be reconciled with any utility function that is normatively appealing given the decision

[9]We refer the interested reader to Choi et al. (2007b) for further details on the testing for consistency with GARP. Choi et al. (2007a) also show that because our subjects make choices in a wide range of budget sets, our data provides a stringent test of utility maximization.

problem at hand. Given the two states in our experiment are equally likely, allocating fewer tokens to the cheaper account ($x_s < x_{s'}$ when $p_s < p_{s'}$) is a violation of monotonicity with respect to FOSD. Violations of FOSD may reasonably be regarded as errors, regardless of risk attitudes—that is, as a failure to recognize that some allocations yield payoff distributions with unambiguously lower returns.[10]

To test whether individual choice behavior satisfies GARP and FOSD, we combine the actual data from the experiment and the mirror-image data and compute the CCEI for this combined dataset.[11] Clearly, always allocating all tokens to one of the accounts generates severe violations of GARP in the combined data set, but the subset of actual data is perfectly consistent. Similarly, any decision to allocate fewer tokens to the cheaper asset will necessarily generate a simple violation of the weak axiom of revealed preference (WARP) involving its mirror-image decision. Polisson et al. (2020) show that when the two states are equally likely (as in our experiment), the CCEI score for the combined dataset—denoted by $e^{**} \leq e^{*} \leq 1$—is a measure of consistency with GARP and FOSD.

What types of risk preferences could give rise to choices consistent with GARP and FOSD? One formulation that encompasses a number of non-EUT models and reduces to EUT would be preferences generated by the RDU (Quiggin, 1982) utility function:

$$U(\tilde{\mathbf{x}}) = \beta_L u(x_L) + \beta_H u(x_H),$$

where β_L, β_H are the decision weights, $\tilde{\mathbf{x}} = (x_L, x_H)$ is the rank-ordered allocation with payoffs $x_L \leq x_H$, and and $u(\cdot)$ is the Bernoulli index. EUT is a special case of RDU when $\beta_L = \beta_H$ (since each state is equiprobable). For any $\beta_L > \beta_H$ the RDU formulation takes the familiar (inverted) s-shaped, interpreted as pessimism—the indifference curves then have a 'kink' at safe allocations where $x_1 = x_2$ (on the 45-degree line). Such allocations will be chosen for a nonnegligible set of price ratios around $p_1 = p_2$, which is inconsistent with EUT (as prices are randomly generated, smooth preferences should give rise to allocations satisfying $x_1 = x_2$ with probability zero).

[10] As noted by Quiggin (1990) and Wakker and Tversky (1993), theories of choice under uncertainty that violated monotonicity with respect to FOSD have been amended to avoid such violations.

[11] The data generated by an individual's choices are $\left\{\left(\bar{\mathbf{x}}^i, \mathbf{x}^i\right)\right\}_{i=1}^{50}$, where $\bar{\mathbf{x}} = (\bar{x}_1^i, \bar{x}_2^i)$ are the endpoints of the budget line. The mirror-image data are obtained by reversing the prices and the associated allocation for each observation $\left\{\left(\bar{x}_2^i, \bar{x}_1^i, x_2^i, x_1^i\right)\right\}_{i=1}^{50}$.

The RDU formula for the rank-ordered allocation $\tilde{\mathbf{x}}$ can be expressed in terms of the probability weighting function w as follows:

$$\beta_L = 1 - w\left(\tfrac{1}{2}\right) \text{ and } \beta_H = w\left(\tfrac{1}{2}\right).$$

That is, the cumulative distribution function of the induced lottery assigns to each monetary payoff the probability of receiving that payoff or anything less. Note that the weighting function w—which is increasing and satisfies $w(0) = 0$ and $w(1) = 1$—transforms the distribution function into decision weights. By definition, the decision weight β_H is equal to $w\left(\frac{1}{2}\right)$ in the case of two states.

For each subject, we estimated the EUT and RDU models using a constant relative risk aversion (CRRA) specification and a constant absolute risk aversion (CARA) specification. For CRRA, we assume $u(\cdot)$ takes the power form $u(x_s) = x_s^{1-\rho}/(1-\rho)$ (with $u(x_s) = \log(x_s)$ if $\rho = 1$), where $\rho \geq 0$ is the Arrow-Pratt measure of relative risk aversion. For CARA, we assume $u(\cdot)$ takes the exponential form $u(x_s) = -e^{-\gamma x_s}$ where $\gamma \geq 0$ is the coefficient of absolute risk aversion. The economic parameter vector is thus $\theta = (w, \rho)$ for CRRA and $\theta = (w, \gamma)$ for CARA. For each subject, we use the specification—CRRA or CARA—that makes more accurate predictions and compare the performance of this specification to the performances of a variety of ML models.

3.4 Machine Learning Models

We consider seven models across three main families of ML models – regularized regressions, tree-based, and neural networks. Each class is commonly used in the ML, and increasingly economics, literatures. We include multiple approaches because there is no declared 'winning' method.[12] For each subject, we consider both the most complete (accurate) ML model within each class, and then additionally the most complete of all eight models considered. We describe our application of the models below, and refer readers to textbooks such as Hastie et al. (2009) and Daumé (2017) for an in-depth treatment that the reader may wish to consult.

[12] As Athey and Imbens (2019) state "[t]here are no formal results that show that, for supervised learning problems, deep learning or neural net methods are uniformly superior to regression trees or random forests, and it appears unlikely that general results for such comparisons will soon be available, if ever[,]"

Regularized regressions Regularized regression, in its simplest form, assumes a linear relationship between outcomes and covariates, whose coefficient is estimated using ordinary least squares with a penalty term. Roughly, the penalty term lets the model "learn" which variables are important, and which to ignore. While including a penalty biases the coefficients, doing so also reduces the chance of overfitting, or "chasing noise." We consider two popular models of regularized regression that add the norm of the coefficient vector as the penalty. The two differ in which norm is implemented as the penalty. First, we consider Lasso (Tibshirani, 1996), which penalizes using the L_1 norm. Formally, estimating relative demand using Lasso generates a mapping $\hat{f}_{Lasso}$:

$$\hat{f}_{Lasso}(\mathcal{B}) = \hat{\beta}^T \mathcal{B},$$

where $\hat{\beta}$ solves

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^{50} (\mathbf{x}^i - \beta^T \mathcal{B}^i)^2 + \lambda \mid\mid \beta \mid\mid_1$$

Second, we consider ridge regression (Hoerl and Kennard, 1970), which penalizes using the L_2 norm. Formally, estimating relative demand using Ridge generates a mapping $\hat{f}_{Ridge}$:

$$\hat{f}_{Ridge}(\mathcal{B}) = \hat{\beta}^T \mathcal{B},$$

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^{50} (\mathbf{x}^i - \beta^T \mathcal{B}^i)^2 + \lambda (\mid\mid \beta \mid\mid_2)^{1/2}$$

The parameter λ affects the degree to which the size of β affects the objective function. If $\lambda = 0$, then the optimization is OLS. We use leave-one-out cross-validation to determine the parameter $\lambda \in [0, 0.2, 0.4, 0.6, 0.8, 1]$. The budget set $\mathcal{B}^i$ is encoded as an intercept $1/p_1$ and the price ratio p_2/p_1. The parameter vector θ for regularized regressions models is a linear coefficient vector.

Tree-based Let t denote one of the possible variables associated with a budget set. Unlike the linear relationship assumed in regularized regression, tree-based models divide the set of budget sets $\boldsymbol{B}$ into subsets (based on the prices and the endowment)

and estimate a model on each of the subsets. This division is done recursively. That is, given some index of observations Z corresponding to data $\{(\mathcal{B}^i, \mathbf{x}^i)\}_{i \in Z}$, the the algorithm considers all further binary partitions that can be represented as separating data based on a variable x being above or below a given threshold k: $\{(\mathcal{B}^i, \mathbf{x}^i)\}_{i \in Z \text{ and } t^i \leq k}$ and $\{(\mathcal{B}^i, \mathbf{x}^i)\}_{i \in Z \text{ and } t^i > k}$. Of these partitions, the selected partition is the (t, k) pair that minimizes error when applying optimal models to each partition.

$$(t^*, k^*) \in \text{argmin}_{(t,k)} \left\{ \sum_{i: i \in Z, t^i \leq k} \ell \left[f_\theta^{\leq}(\mathcal{B}^i), \mathbf{x}^i \right] + \sum_{i: i \in Z, t^i > k} \ell \left[f_\theta^{>}(\mathcal{B}^i), \mathbf{x}^i \right] \right\},$$

where

$$f_\theta^{\leq} = \text{argmin}_{f \in \mathcal{F}_\Theta} \sum_{i: i \in Z, t^i \leq k} \ell(f(\mathcal{B}^i, \mathbf{x}^i))$$

and

$$f_\theta^{>} = \text{argmin}_{f \in \mathcal{F}_\Theta} \sum_{i: i \in Z, t^i > k} \ell(f(\mathcal{B}^i, \mathbf{x}^i)).$$

The process is then reapplied for the two subsets of the resulting partition, and so on. This partitioning process generates both the (locally) best partition of budget sets and the (locally) best model estimate for the partition. In aggregate, the algorithm returns a piecewise demand function. To predict the relative demand of some budget set $\mathcal{B}^i$, first the subset containing $\mathcal{B}^i$ determines which model to use. Then, evaluating that model determines the demand.

This partitioning process, if allowed to continue without restraint, would end with each data point in its own partition, with perfect within-sample prediction. To prevent such overfitting, we limit the decision trees in two simple ways. First, we set a minimum number of observations per partition. This prevents the algorithm from splitting a partition if doing so would result in an insufficiently large sample size. Second, we limit the "depth", or number of partitions away from $\boldsymbol{B}$, of a tree. These limits are determined endogenously for each subject by performing 3-fold cross validation. In this procedure, data is randomly split into three equally sized subsamples. We choose the maximum depth to search over 2, 4, 6, and 8; we choose the minimum observations per partition to search over 2, 4, 6, 8, and 10.

The standard decision tree model, denoted Mean, takes the sample mean token share x of each subset. We use Mean as well as three extensions. The first extension, known more broadly as model trees (Quinlan et al., 1992), changes the estimated model from a sample mean to a linear regression (Linear). Mean is nested in Linear. The second extension, support vector regression trees, instead uses a support vector regression of each subset. Support vector regression attempts to find the flattest demand mapping possible such that the token share predictions are accurate up to some $\epsilon \geq 0$ (see Smola and Schölkopf, 2004). The last tree-based model, the random forest model (RF), averages the decision rules of multiple standard decision trees. Each tree is given a bootstrapped data set, and is generally seen as an improvement over singular decision trees (Breiman, 2001). In addition to limits on depth and minimum sample size, RF regulates the number of trees, which we choose to be 10, 50, and 100 trees. Because each tree not trained on the original data set, there is no nesting and thus no restrictiveness or completeness guarantees between RF and the other tree-based models. Additionally, since trees are inherently nonparametric, they cannot be easily described by a parameter vector θ.

Neural networks Neural networks, specifically a multilayer perceptron, transform budget sets into relative demand predictions by nonlinear regression, whose functional form assumes a series of nested transformations. In our setup, the transformation takes two parts. First, a budget set $\mathcal{B}$ undergoes an affine transformation $W^{(0)}\mathcal{B}+b^{(0)}$, where $W^{(0)}$ and $b^{(0)}$ are a matrix and vector of size $n_0 \times 2$ and $n_0 \times 1$, respectively. The dimension n_0 is prespecified by the analyst. Second, the affine transformation is again transformed by a function $\sigma^{(0)} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_0}$ to obtain a new vector $\mathcal{B}^{(1)} = \sigma^{(0)}(W^{(0)}\mathcal{B}+ b^{(0)})$. The function σ is also prespecified by the analyst. The resulting vector, $\mathcal{B}^{(1)}$, is referred to as a "hidden layer". It is then used as the input to generate another hidden layer, $\mathcal{B}^{(2)} = \sigma^{(1)}(W^{(1)}\mathcal{B} + b^{(1)})$, using a new affine transformation defined by $\underset{n_1 \times n_0}{W^{(1)}}$ and $\underset{n_1 \times 1}{b^{(1)}}$ as well as transformation by $\sigma^{(1)}$. This process continues for the number of hidden layers prespecified by the analyst. The final affine transformation results in a scalar value that can be interpreted as the estimated relative demand.

For a multilayer perceptron, the parameter values $W^{(i)}$ and $b^{(i)}$ are estimated, while the analyst has the freedom to choose the number of layers, the dimensions of each layer, the $\sigma^{(i)}$ functions, and a number of parameters associated with the estimation of $W^{(i)}$ and $b^{(i)}$.

We use the layer count, layer dimension, and $\sigma^{(i)}$ values from Hsieh et al. (2023). $\sigma^{(i)}$ are all chosen to be the same component-wise maximum function $\sigma(x) = \max(0, x)$. This function, the rectified linear unit ("ReLU") function, keeps all positive components of a vector, and sets all negative components to zero. We use 3-fold cross-validation to simultaneously determine the individual-best layer count and layer dimension. We search over all combinations of $\{1, 2, 3\}$ hidden layers, as well as all combinations of $\{15, 20, 25\}$ for the size of each layer, for a total of 39 "architectures" investigated.

We use the L-BFGS algorithm (Liu and Nocedal, 1989) to estimate $W^{(i)}$ and $b^{(i)}$. This is a standard optimization algorithm that uses first and second order information to iteratively update estimates of parameter values. This algorithm is generally not feasible to compute for larger models and larger data sets, but is applicable in our setting with 50 observations per subject. For a full treatment, see Bottou et al. (2018) and Sun et al. (2019). The estimation objective function to be minimized is mean squared error, which is the same objective function used to evaluate all models (through completeness and restrictiveness). For example, given a network of 2 hidden layers each with dimension 15, the objective function is:

$$\min_{\substack{W^{(0)}_{15\times 2}, W^{(1)}_{15\times 15}, W^{(2)}_{1\times 15}, b^{(0)}_{15\times 1}, b^{(1)}_{15\times 1}, b^{(2)}_{1\times 1}}} \sum_{i=1}^{50} \ell(f(W,b), x^i) = \\ = \left[x^i - W^{(2)}\sigma\left(W^{(1)}\sigma\left(W^{(0)}\mathcal{B}^i + b^{(0)}\right) + b^{(1)}\right) - b^{(2)}\right]^2$$

4 Results

Table 1 provides a population-level summary of our results, complementing the information provided in Figure 1 above. The left column of Table 1 reports the average completeness of each model, as well the 95% confidence interval for average completeness, and the next column reports the win rate of EUT against each model (that is, the fraction of subjects for whom EUT is more complete). The next two blocks of four columns report the win rate of EUT against each model and its absolute completeness difference by quartiles of the consistency score with GARP and FOSD. The right column reports the restrictiveness of each model. Panel A of Table 1 reports the results for the three *families* of ML models—regularized regressions, tree-based,

and neural networks. For regularized regressions and tree-based models, we report restrictiveness as weighted averages of the most complete model in the class for each subject. Panels B and C of Table 1 report the results for each regularized regression and tree-based model, respectively.

[Table 1 here]

Three main insights arise from Panel A of Table 1 about the prediction accuracy (completeness) and model flexibility (restrictiveness) of EUT as compared to that of the families of ML models. Similar insights arise from Panels B and C when comparing the economic model to each regularized regression and tree-based model:

- First, the completeness of EUT is comparable to the completeness of the tree-based models (achieving 89.3% and 89.1% of the feasible reduction in prediction error, respectively), but it is significantly more complete than regularized regression models and neural networks (achieving completeness of only 79.5% and 71.6%, respectively). Furthermore, EUT's completeness win rate increases from 69.7% against tree-based models to 88.5% against regularized regression models and to 94.2% against neural networks.

- Second, the win rate of EUT almost always increases by consistency quartiles against all three families of ML models, as well as its relative improvement over regularized regression models and neural networks. Perhaps as expected, the predictive accuracy of EUT is improved compared to the accuracy of ML models when individual choices more closely satisfy the axioms on which the economic model is based.

- Third, while EUT does not achieve a large improvement in completeness compared to tree-based models, it is substantially more restrictive (18.6% compared to only 10.6%). Moreover, the restrictiveness of EUT is comparable to the restrictiveness of the regularized regression models and neural networks (achieving restrictiveness of 20.7% and 14.4%, respectively), but these ML models are significantly less complete than EUT.

We can see the comparison of completeness between EUT and the most complete ML model in greater detail in the four panels of Figure 2 below corresponding to the

quartiles of the consistency of the individual-level data with GARP and FOSD. For each subject, the horizontal axis in each panel shows the completeness of EUT and the vertical horizontal axis shows the completeness of the best ML model. Each axis also provides a marginal kernel density estimate of completeness scores approximated using a Gaussian kernel.

[Figure 2 here]

We first observe that there are relatively few extreme differences in completeness, as indicated by the absence of observations in the upper left and lower right corners of each panel. However, there are a few observations high above the diagonal in the bottom consistency quartile. Additionally, we note a monotonic shift towards the upper right corner by consistency quartiles, indicating greater completeness of both models as individual choices become more consistent. The fraction of observations below the diagonal (subjects for which EUT is the most complete model) weakly increases by consistency quartile, and the distribution of completeness is higher for EUT in all panels. Finally, we note a complementarity between ML models—of the 334 subjects for whom the most complete ML model is more complete than EUT, 245 (73.3%) have EUT as the second most complete model, above the other two classes of ML models.

Recall that the RDU model reduces to EUT when $\beta_L = \beta_H$. RDU is therefore less restrictive than EUT (18.6% compared to 16.6%) and it is also only moderately less restrictive than the regularized regression models (16.6% compared to 20.7%). Table 2 below provides a population-level summary of our results comparing the economic models, EUT and RDU, in the same format as Table 1. Panel A of Table 2 reports the results taking a weighted average of the most complete $u(\cdot)$ specification for each subject, CRRA or CARA. Panels B and C of table 2 report the results assuming $u(\cdot)$ takes the CRRA and CARA specifications, respectively. The main insights that arise from Table 2 are that the average completeness of EUT is the same as the completeness of RDU but it is more restrictive. The absolute improvement of RDU over EUT is essentially zero under both CARA and CRRA in all consistency quartiles. In Online Appendix, we present near-identical results for Figure 1, Table 1, and Figure 2 with RDU instead of EUT.

[Table 2 here]

Finally, although the number of individual decisions in our experiments is much higher than typically seen in the experimental literature—providing us with a rich dataset of individual decisions across a wide range of budget lines for a robust test—we would still want to compare the completeness of EUT versus ML if the number of individual decisions is significantly larger than what a human subject can handle in a single experimental session. To this end, we generate a random sample of hypothetical subjects who implement EUT with the power Bernoulli utility function $u(x_s) = x_s^{1-\rho}/(1-\rho)$ with an idiosyncratic preference shock that has a logistic distribution (so the likelihood of error is a decreasing function of the utility cost of an error):

$$\Pr(\mathbf{x}) = \frac{e^{\xi \cdot U(\mathbf{x}^*)}}{\int\limits_{\mathbf{x}:\mathbf{p}\cdot\mathbf{x}=1} e^{\xi \cdot U(\mathbf{x})}},$$

where the precision parameter ξ reflects sensitivity to differences in utility. The choice of portfolio becomes purely random as $\xi \to 0$, whereas the probability of the portfolio yielding the highest expected utility approaches one as $\xi \to \infty$.

We generated samples of hypothetical subjects with $\rho = 1/2$, which is in the range of our human subjects estimated risk aversion, and four levels of $\xi = 0, 0.25, 1, 10$. Each of the hypothetical subject makes 1,000 choices from randomly generated budget sets in the same way as the human subjects do. We compare the completeness of EUT and the most complete ML model for the hypothetical subjects in the four panels of Figure 3 below—corresponding to the four levels of ξ. For each subject, the horizontal axis in each panel shows the completeness of EUT and the vertical horizontal axis shows the completeness of the best ML model. Each axis also provides a marginal kernel density estimate of completeness scores approximated using a Gaussian kernel. The scatterplots clearly show that ML does not outperform EUT, even when the number of individual decisions is very large, across all levels of error.

[Figure 3 here]

5 Conclusion

We use graphical representations of budget lines over bundles of state-contingent commodities, enabling us to collect a rich individual-level dataset. Our analysis starts with applying revealed preference tests to check if the observed choices align

with the axioms of economic theory—ordering (completeness and transitivity) and monotonicity (with respect to FOSD). We then compare the completeness of EUT versus non-EUT and various ML models at the individual level. Our main finding is that both the standard EUT model and the RDU model outperform all ML models, with a wider margin as individual choices become more consistent with an underlying (monotonic) preference ordering. We view this as a 'victory' for the economic models, particularly EUT, as it is nested within RDU and thus more restrictive.

The experimental and analytical techniques developed in this paper lay the groundwork for studying decision-making under risk and uncertainty in more complex scenarios. In a concurrent paper by Ellis et al. (2024), we analyze similar experimental data involving choices under risk with three states and three associated securities, where the probabilities of all states are objectively known and equal. For three states, both EUT and prominent non-EUT models impose specific and quite stringent restrictions on the utility function's structure, leading to empirically testable predictions about observed behavior. Additionally, Ellis et al. (2024) examine the completeness of models of choice under ambiguity and compare them to the completeness of ML models. To differentiate the effects of risk (known probabilities) and ambiguity (unknown probabilities), one state has an objectively known probability, while the probabilities of the other two states are unknown. The analysis with three states is not a straightforward extension of the two-state case and is computationally intensive for large datasets like ours.

An important advantage of the methods and analyses presented in this paper is their adaptability to different decision domains. While the experiment reported here focused on choices involving risk, in ongoing work we explore intertemporal choice. The long-standing interest in intertemporal choice has been further fueled by recent evidence of non-constant time discounting and a deeper understanding of its theoretical implications. Additionally, in other ongoing work, we generate and analyze new experimental data to understand preferences towards risk and time using generally representative subject pools (as in Choi et al., 2014). This work aims to link our high-quality experimental data to survey data about economic behavior in real-world settings—using both economic models and ML models—to improve predictions and evaluate alternative theories about important economic decisions.

We view this as the beginning of a much broader agenda that builds on the experimental methodology and analytical techniques developed in this paper. However,

examining behavior in other, more complex settings will naturally require further experimental data as well as new theoretical and analytical techniques.

References

Afriat, S. N. (1967): "The Construction of Utility Functions From Expenditure Data," *International Economic Review*, 8, 67–77.

——— (1972): "Efficiency Estimation of Production Functions," *International Economic Review*, 568–598.

Almog, D., R. Gauriot, L. Page, and D. Martin (2024): "AI Oversight and Human Mistakes: Evidence from Centre Court," *arXiv preprint arXiv:2401.16754*.

Andrews, I., D. Fudenberg, L. Lei, A. Liang, and C. Wu (2024): "The Transfer Performance of Economic Models," *arXiv preprint arXiv:2202.04796*.

Athey, S. and G. W. Imbens (2019): "The Construction of Utility Functions From Expenditure Data," *Annual Review of Economics*, 11, 685–725.

Azrieli, Y., C. P. Chambers, and P. J. Healy (2018): "Incentives in Experiments: A Theoretical Analysis," *Journal of Political Economy*, 126, 1472–1503.

Becker, G. M., M. H. DeGroot, and J. Marschak (1964): "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, 9, 226–232.

Bottou, L., F. E. Curtis, and J. Nocedal (2018): "Optimization Methods for Large-Scale Machine Learning," *SIAM Review*, 60, 223–311.

Breiman, L. (2001): "Random Forests," *Machine Learning*, 45, 5–32.

Brown, A. L. and P. J. Healy (2018): "Separated Decisions," *European Economic Review*, 101, 20–34.

Brunnermeier, M. K., R. Lamba, and C. Segura-Rodriguez (2023): "Inverse Selection," *Available at SSRN 3584331*.

Brynjolfsson, E., D. Li, and L. R. Raymond (2023): "Generative AI at Work," Tech. rep., National Bureau of Economic Research.

Cappelen, A. W., S. Kariv, E. Ø. Sørensen, and B. Tungodden (2023): "The Development Gap in Economic Rationality of Future Elites," *Games and Economic Behavior*, 142, 866–878.

Charness, G., B. Jabarian, and J. A. List (2023): "Generation Next: Experimentation with AI," Tech. rep., National Bureau of Economic Research.

Chen, Y., T. X. Liu, Y. Shan, and S. Zhong (2023): "The Emergence of Economic Rationality of GPT," *Proceedings of the National Academy of Sciences*, 120, e2316205120.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68.

Choi, S., R. Fisman, D. Gale, and S. Kariv (2007a): "Consistency and Heterogeneity of Individual Behavior Under Uncertainty," *American Economic Review*, 97, 1921–1938.

Choi, S., R. Fisman, D. M. Gale, and S. Kariv (2007b): "Revealing Preferences Graphically: An Old Method Gets a New Tool Kit," *American Economic Review*, 97, 153–158.

Choi, S., S. Kariv, W. Müller, and D. Silverman (2014): "Who is (more) rational?" *American Economic Review*, 104, 1518–50.

Clithero, J. a., J. J. Lee, and J. Tasoff (2023): "Supervised Machine Learning for Eliciting Individual Demand," *American Economic Journal: Microeconomics*, 15, 146–182.

Daumé, H. (2017): "A Course in Machine Learning: Hal Daumé III," .

Deaton, A. and J. Muellbauer (1980): "An Almost Ideal Demand System," *American Economic Review*, 70, 312–326.

Diecidue, E. and P. P. Wakker (2001): "On the Intuition of Rank-Dependent Utility," *Journal of Risk and Uncertainty*, 23, 281–298.

Ellis, K., S. Kariv, and E. Ozbay (2024): "Predicting and Understanding Individual-Level Choice Under Uncertainty," *Working Paper*.

Fehr, E., T. Epper, and J. Senn (2023): "The Fundamental Properties, Stability and Predictive Power of Distributional Preferences," *CESifo Working Paper*.

Fernández-Villaverde, J., S. Hurtado, and G. Nuno (2023): "Financial Frictions and the Wealth Distribution," *Econometrica*, 91, 869–901.

Fudenberg, D., W. Gao, and A. Liang (2023): "How Flexible Is That Functional Form? Quantifying the Restrictiveness of Theories," *The Review of Economics and Statistics*, Forthcoming.

Fudenberg, D. and G. Karreskog Rehbinder (2024): "Predicting Cooperation With Learning Models," *American Economic Journal: Microeconomics*, 16, 1–32.

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2022): "Measuring the Completeness of Economic Models," *Journal of Political Economy*, 130, 956–990.

Fudenberg, D. and A. Liang (2019): "Predicting and Understanding Initial Play," *American Economic Review*, 109, 4112–41.

Fudenberg, D. and I. Puri (2022a): "Evaluating and Extending Theories of Choice Under Risk," *Working Paper*.

——— (2022b): "Simplicity and Probability Weighting in Choice Under Risk," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 112, 421–425.

Gul, F. (1991): "A Theory of Disappointment Aversion," *Econometrica*, 667–686.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2, Springer.

Hoerl, A. E. and R. W. Kennard (1970): "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Horton, J. J. (2023): "Large Language Models as Simulated Economic Agents: What Can We Learn From Homo Silicus?" Tech. rep., National Bureau of Economic Research.

Hsieh, S.-L., S. Ke, Z. Wang, and C. Zhao (2023): "A Logit Neural-Network Utility Model," *Working Paper*.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, 47, 363–391.

Kim, J., M. Kovach, K.-M. Lee, E. Shin, and H. Tzavellas (2024): "Learning to be Homo Economicus: Can an LLM Learn Preferences from Choice," *arXiv preprint arXiv:2401.07345*.

Kobayashi, S. J. and A. Lucia (2023): "Robust Estimation of Risk Preferences," *Working Paper*.

Liu, D. C. and J. Nocedal (1989): "On the Limited Memory BFGS Method for Large Scale Optimization," *Mathematical Programming*, 45, 503–528.

Ludwig, J. and S. Mullainathan (2024): "Machine Learning as a Tool for Hypothesis Generation," *The Quarterly Journal of Economics*, 139, 751–827.

Machina, M. J. (1994): "Review of Generalized Expected Utility Theory: The Rank-Dependent Model," *Journal of Economic Literature*, 32, 1237–1238.

Mullainathan, S. and A. Rambachan (2024): "From Predictive Algorithms to Automatic Generation of Anomalies," Tech. rep., National Bureau of Economic Research.

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

Nishimura, H., E. A. Ok, and J. K.-H. Quah (2017): "A Comprehensive Approach to Revealed Preference Theory," *American Economic Review*, 107, 1239–63.

Peterson, J. C., D. D. Bourgin, M. Agrawal, D. Reichman, and T. L. Griffiths (2021): "Using Large-Scale Experiments and Machine Learning to Discover Theories of Human Decision-Making," *Science*, 372, 1209–1214.

Peysakhovich, A. and J. Naecker (2017): "Using Methods From Machine Learning to Evaluate Behavioral Models of Choice Under Risk and Ambiguity," *Journal of Economic Behavior & Organization*, 133, 373–384.

Polisson, M., J. K.-H. Quah, and L. Renou (2020): "Revealed Preferences over Risk and Uncertainty," *American Economic Review*, 110, 1782–1820.

Puri, I. (2018): "Preference for Simplicity," *Available at SSRN 3253494*.

Quiggin, J. (1982): "A Theory of Anticipated Utility," *Journal of Economic Behavior & Organization*, 3, 323–343.

——— (1990): "Stochastic Dominance in Regret Theory," *The Review of Economic Studies*, 57, 503–511.

Quinlan, J. R. et al. (1992): "Learning with Continuous Classes," in *5th Australian Joint Conference on Artificial Intelligence*, World Scientific, vol. 92, 343–348.

Smola, A. J. and B. Schölkopf (2004): "A Tutorial on Support Vector Regression," *Statistics and Computing*, 14, 199–222.

Starmer, C. (2000): "Developments in Non-expected Utility Theory: The Hunt for a Descriptive Theory of Choice Under Risk," *Journal of Economic Literature*, 38, 332–382.

Sun, S., Z. Cao, H. Zhu, and J. Zhao (2019): "A Survey of Optimization Methods From a Machine Learning Perspective," *IEEE Transactions on Cybernetics*, 50, 3668–3681.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

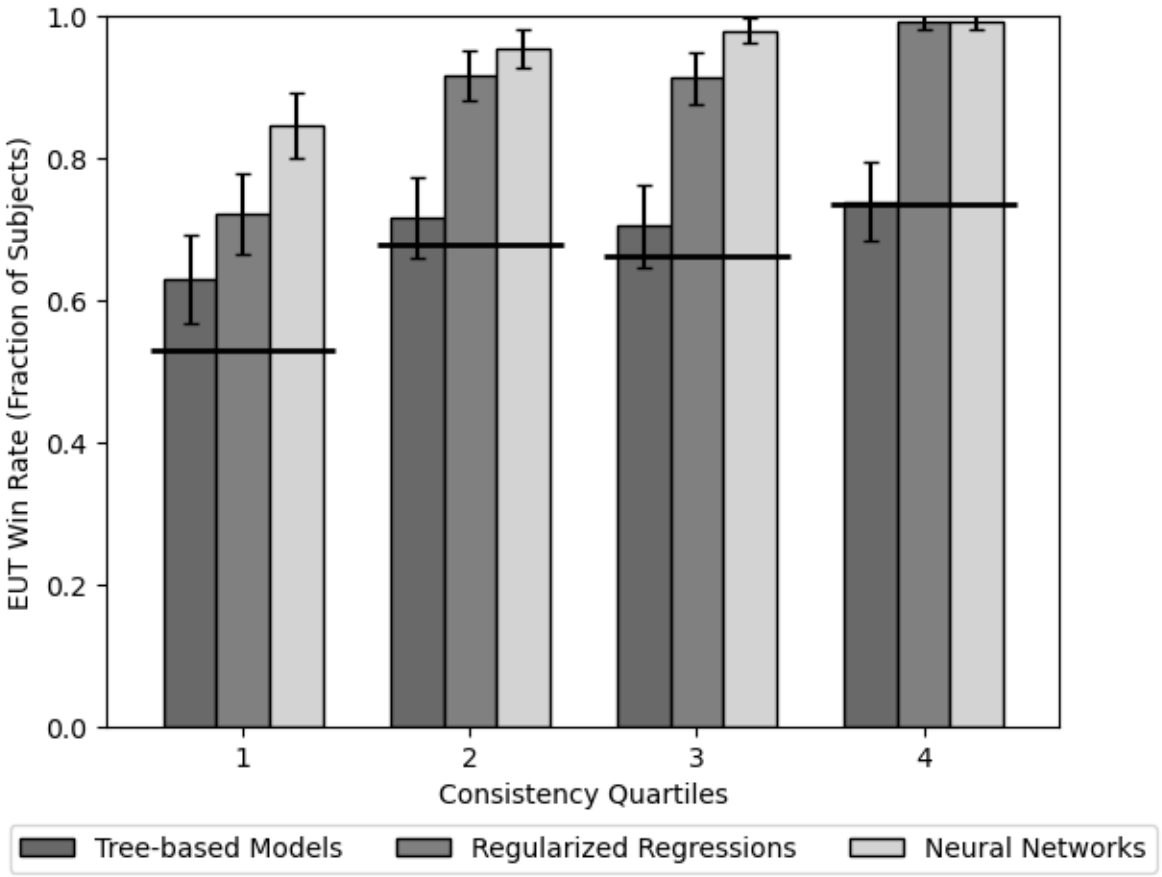
Varian, H. R. (1982): "The Nonparametric Approach to Demand Analysis," *Econometrica*, 945–973.

——— (1983): "Non-parametric Tests of Consumer Behaviour," *The Review of Economic Studies*, 50, 99–110.

Wakker, P. and A. Tversky (1993): "An Axiomatization of Cumulative Prospect Theory," *Journal of Risk and Uncertainty*, 7, 147–175.

Zame, W. R., B. Tungodden, E. Ø. Sørensen, S. Kariv, and A. W. Cappelen (2024): "Linking Social and Personal Preferences: Theory and Experiment," *Working Paper*.

Figure 1: EUT win rate over ML by quartiles of consistency scores with GARP and FOSD



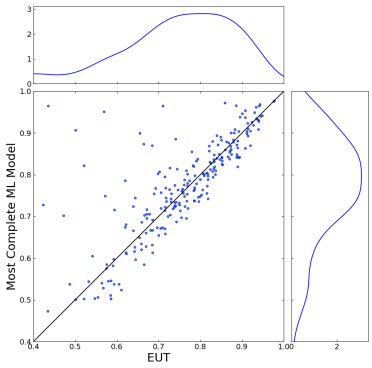
The fraction of subjects for whom EUT is more complete than the best regularized regression, tree-based, and neural network models, as well as the overall best ML model (indicated by black horizontal lines). The x-axis groups subjects by quartiles of consistency scores with GARP and FOSD, following the methods of Nishimura et al. (2017) and Polisson et al. (2020). This score measures the amount by which each budget constraint must be relaxed in order to remove all violations of GARP and FOSD and it is bounded between 0 and 1. A score closer to 1 indicates stronger consistency with GARP and FOSD. The quartile ranges are $[0, 0.83)$, $[0.83, 0.95)$, $[0.95, 0.99)$, and $[0.99, 1]$.

Table 1: The completeness and restrictiveness of EUT versus ML models

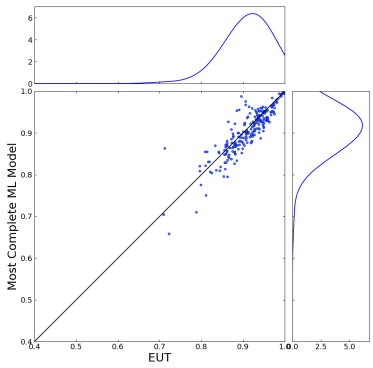
	Average	EUT's win rate	EUT's win rate against ML by e^{**} quartiles				Absolute completeness difference between EUT and ML by e^{**} quartiles				
Panel A: Model Classes	Completeness	against ML	1st	2nd	3rd	4th	1st	2nd	3rd	4th	Restrictiveness
EUT	89.3% [88.4%, 90.0%]	-	-	-	-	-	-	-	-	-	18.6%
Regularized Regressions	79.5% [77.8%, 80.5%]	88.5%	72.1%	91.6%	91.3%	99.2%	3.7%	7.8%	9.7%	17.8%	20.7%
Tree-based Models	89.1% [88.3%, 89.9%]	69.7%	62.9%	71.5%	70.4%	73.8%	-1.4%	0.9%	0.5%	0.6%	10.6%
Neural Networks	71.6% [68.8%, 73.7%]	94.2%	84.6%	95.4%	97.9%	99.2%	9.3%	14.6%	16.6%	30.5%	14.4%
Panel B: Regularized regressions											
Lasso	75.9% [74.2%, 76.9%]	92.2%	80.0%	95.0%	94.2%	99.6%	7.0%	11.8%	13.8%	21.1%	20.7%
OLS	70.2% [57.7%, 74.6%]	90.3%	75.8%	92.5%	93.8%	99.2%	11.2%	10.6%	15.8%	38.8%	20.7%
Ridge	70.6% [58.2%, 75.1%]	90.3%	75.8%	92.5%	93.8%	99.2%	11.1%	10.4%	15.4%	37.9%	20.7%
Panel C: Tree-based models											
Mean	86.6% [85.6%, 87.4%]	84.8%	79.2%	89.1%	86.3%	84.8%	3.0%	3.7%	2.2%	1.9%	12.4%
Linear	82.9% [81.7%, 84.0%]	87.1%	83.3%	87.4%	87.1%	90.7%	12.4%	6.1%	3.5%	3.4%	5.4%
SVR	85.7% [84.8%, 86.6%]	89.1%	80.8%	88.7%	92.5%	94.5%	4.1%	4.1%	2.7%	3.3%	10.7%
RF	88.0% [87.2%, 88.8%]	80.9%	73.3%	81.2%	82.9%	86.1%	0.5%	1.7%	1.2%	1.5%	11.9%

The left column reports the average completeness of each model, as well the 95% confidence interval for average completeness, and the next column reports the win rate of EUT against each model (that is, the fraction of subjects for whom EUT is more complete). The next two blocks of four columns report the win rate of EUT against each model and its absolute completeness difference by quartiles of the consistency score with GARP and FOSD. The right column reports the restrictiveness of each model. Panel A reports the results for EUT and the three *families* of ML models—regularized regressions, tree-based, and neural networks. For regularized regressions and tree-based models, we report restrictiveness as weighted averages of the most complete model in the class for each subject. Panels B and C report the results for each regularized regression and tree-based model, respectively.

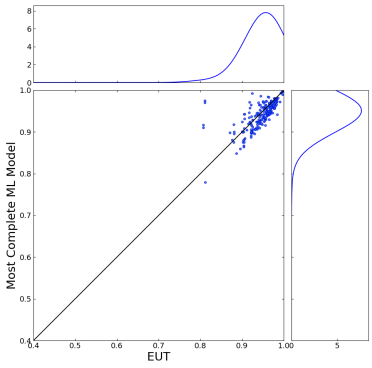
Figure 2: The individual-level completeness of EUT versus the most complete ML model by e^{**} quartile.



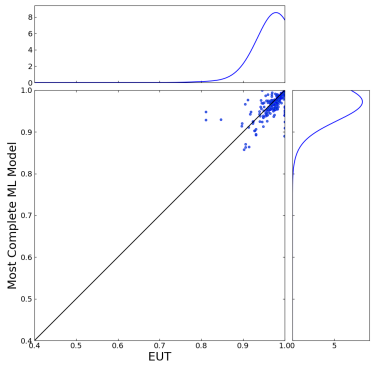
(a) e^{**} quartile 1



(b) e^{**} quartile 2



(c) e^{**} quartile 3



(d) e^{**} quartile 4

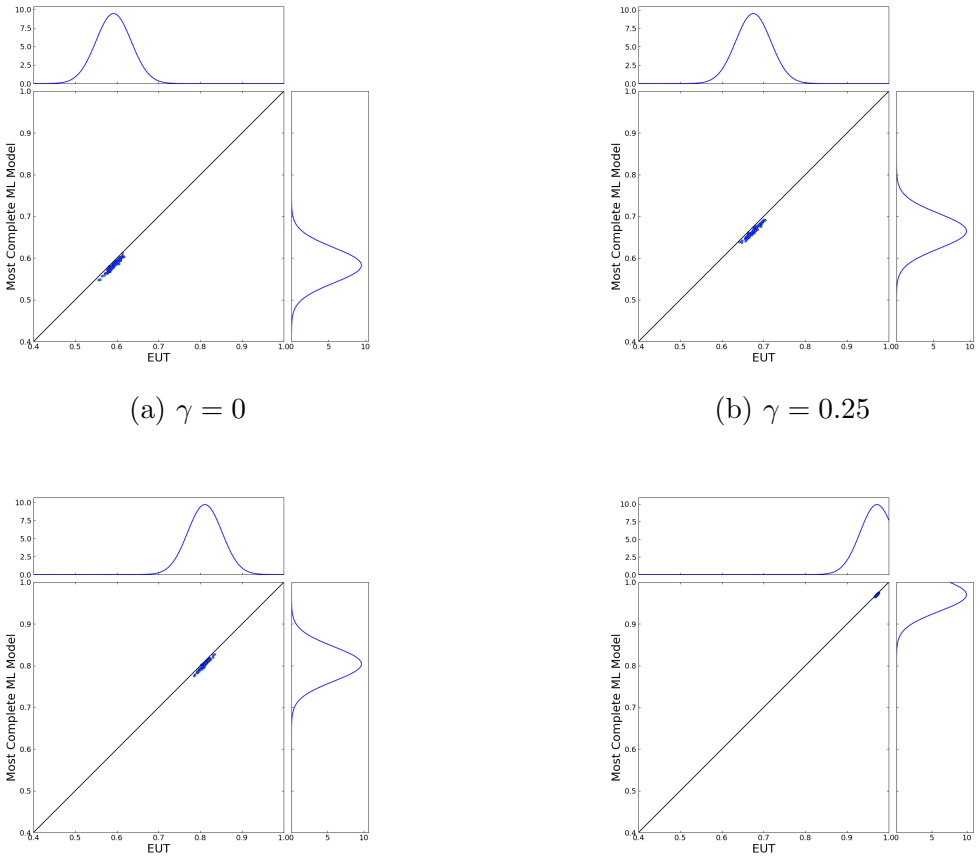
The four panels plot the completeness scores of all subjects for EUT and the best ML model. Panels refer to the quartile of consistency score; Panel (a) plots the subjects in the lowest quartile of e^{**}, Panel (b) the second quartile of e^{**}, and so on. The quartile ranges are $[0, 0.83)$, $[0.83, 0.95)$, $[0.95, 0.99)$, and $[0.99, 1]$. Each plotted point represents a subject. The horizontal axes are the completeness of EUT, and the vertical axes show the completeness of the best ML model. Each axis also provides a marginal kernel density estimate of completeness scores approximated using a Gaussian kernel.

Table 2: The completeness and restrictiveness of EUT versus RDU

	Average	EUT win rate	EUT's win rate against RDU by e^{**} quartiles				Absolute completeness difference between EUT and RDU by e^{**} quartiles				
Panel A: EUT and RDU	completeness	against RDU	1st	2nd	3rd	4th	1st	2nd	3rd	4th	Restrictiveness
EUT	89.3% [88.4%, 90.0%]										18.6%
RDU	89.2% [88.3%, 89.9%]	60.6%	68.3%	66.5%	53.8%	53.6%	0.6%	0.2%	-0.2%	-0.2%	16.6%
Panel B: CRRA Only											
EUT CRRA	88.8% [88.0%, 89.6%]										17.9%
RDU CRRA	88.8% [87.9%, 89.6%]	53.0%	65.4%	59.0%	43.3%	44.3%	0.7%	0.2%	-0.4%	-0.3%	16.3%
Panel C: CARA Only											
EUT CARA	88.6% [87.8%, 89.4%]										19.3%
RDU CARA	88.5% [87.6%, 89.3%]	58.3%	65.4%	64.9%	56.3%	46.4%	0.5%	0.2%	-0.2%	-0.2%	16.9%

The left column reports the average completeness of each model, as well the 95% confidence interval for average completeness, and the next column reports the win rate of EUT against RDU. The next two blocks of four columns report the win rate of EUT against RDU and its absolute completeness difference by quartiles of the consistency score with GARP and FOSD. The right column reports the restrictiveness of each model. Panel A of reports the results taking a weighted average of the most complete $u(\cdot)$ specification for each subject, CRRA or CARA. Panels B and C report the results assuming $u(\cdot)$ takes the CRRA and CARA specifications, respectively.

Figure 3: The individual-level completeness of EUT versus the most complete ML model on noisy subjects with 1,000 choices.



(a) $\gamma = 0$ (b) $\gamma = 0.25$

(c) $\gamma = 1$ (d) $\gamma = 10$

The four panels plot the completeness scores of all subjects for EUT and the best ML model. Panels refer to the simulated noise levels. Each plotted point represents a simulation with 1,000 observations. The horizontal axes are the completeness of EUT, and the vertical axes show the completeness of the best ML model. Each axis also provides a marginal kernel density estimate of completeness scores approximated using a Gaussian kernel.