Predicting and Understanding Individual-Level Choice Under Risk*

Keaton Ellis, Shachar Kariv and Erkut Ozbay†

December 7, 2025

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 line. 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.

JEL Classification Numbers: C63, C91, D81.

Keywords: machine learning, revealed preference, risk preferences, expected utility, non-expected utility, completeness, restrictiveness, experiments.

*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, Matthew Polisson, and Anna Vakarova for helpful conversations. The paper has also benefited from suggestions by the participants of D-TEA 2024, RUD 2024, WEAI 2024, MLESC24, ESIF-AIML2024, 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: Monash University (keaton.ellis@monash.edu); Kariv: University of California, Berkeley (kariv@berkeley.edu); Ozbay: University of Maryland (ozbay@umd.edu).

1 Introduction

This paper explores the potential of machine learning (ML) models to predict *individual-level* choices under risk, and compares their performance to standard economic models, including Expected Utility Theory (EUT). 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 with the axioms of economic models at the individual level does not imply that aggregate data are consistent. In fact, the considerable heterogeneity in subjects' behavior 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).

Specifically, 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 line. 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 dataset. 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.

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 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* as defined by Fudenberg et al. (2022, 2025). 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 discerns 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 dataset 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 lines.

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 First-Order Stochastic Dominance (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.[1]

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 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. We therefore prioritize EUT as our representative economic model to compare to ML models.

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.

[1] 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 are the leading contender. Also note that Rank-Dependent Utility (RDU) and Disappointment Aversion (Gul, 1991) coincide in form when there are two equally likely states. See Diecidue and Wakker (2001) for a comprehensive discussion.

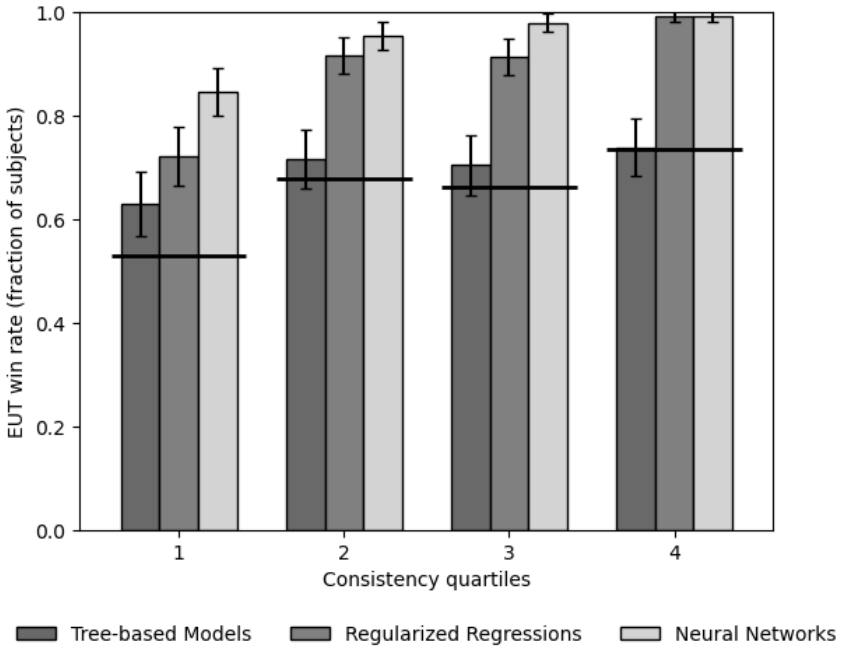
Figure 1 depicts our main result. The horizontal axis presents quartiles over the distribution of subjects' consistency scores with GARP and FOSD. The vertical axis indicates the fraction of subjects for whom EUT is *more* complete than the *most* complete ML model within each class, as well as *more* 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 consistent with 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. The very few cases where ML outperforms the economic models involve subjects whose choices are inconsistent with GARP and/or FOSD, but in a systematic way. For instance, ML is significantly more complete for the two subjects—out of nearly a thousand—who always selected the same budget line endpoint. Their behavior satisfies GARP but severely violates FOSD, as they allocated all tokens to the more expensive security in about half of the rounds.

In contrast to our individual-level results, the best ML model outperforms the economic models when we pool data across all subjects—consistent with previous literature comparing model completeness at the aggregate level. The reason is that individual consistency with GARP and FOSD does not yield aggregate consistency, given heterogeneity in behavior. In addition, multiple levels of heterogeneity may exist between the individual-level analysis used in this paper and the aggregate-level analyses common in the literature. To this end, we conduct two clustering exercises, generating either five or ten clusters and then calculating completeness for each cluster–model pair. As analysis levels become coarser, ML models gain relative completeness over EUT, and although RDU surpasses EUT in aggregate data, it does not outperform EUT at the clustered levels.

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

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



The histograms show the fraction of subjects for whom EUT is more complete than the best regularized regression, tree-based, and neural network models. Black horizontal lines indicate the fraction of subjects for whom EUT is more complete than all ML models. The horizontal axis groups subjects into quartiles of consistency scores with GARP and FOSD. This score, bounded between 0 and 1, measures the extent to which budget constraints must be relaxed to eliminate all violations; scores closer to 1 indicate stronger consistency. The quartile ranges are [0, 0.83), [0.83, 0.95), [0.95, 0.99), and [0.99, 1]. Error bars represent 95% binomial proportion confidence intervals.

interpret our results as a 'victory' for economic models, particularly EUT, which is foundational to much of economics.

Finally, we note that while 50 observations per subject are relatively large by experimental economics standards, they may be considered small in the context of ML applications. To address this, our final analysis calculates completeness scores for simulated subjects making up to 1,000 decisions under EUT and RDU models with varying levels of noise. When the data-generating process follows EUT, EUT is at least as complete as both RDU and the best-performing ML model across all data sizes and noise levels. When the data-generating process follows RDU, RDU is the most complete model when noise is low, as expected. However, EUT becomes more complete when data size is small and noise is high. Importantly, EUT remains more

complete than the best-performing ML model on RDU-generated data except when data size is large and noise is low.

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. All technical details and the experimental instructions are available in the Online Appendix.

2 Related literature

Our paper contributes to the body of work that seeks to use ML techniques to enhance economic models – theoretical and empirical. We will not attempt to review this large and growing literature or offer a full overview of its applications.[2] Instead, we focus attention on the recent papers that are particularly relevant to our study.

Peysakhovich and Naecker (2017) were among the first to apply ML as a predictive benchmark, comparing the performances of EUT and prominent non-EUT alternatives to the performance of ML 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 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.[3]

[2]The introduction of ML has fundamentally transformed economics with examples spanning labor economics (Brynjolfsson et al., 2025), macroeconomics (Fernández-Villaverde et al., 2023), econometrics (Chernozhukov et al., 2018), and experimental economics (Horton, 2023). Emerging subfields of human–artificial intelligence (AI) interaction can be seen in behavioral and experimental economics (Charness et al., 2025; Almog et al., 2024), mechanism design (Brunnermeier et al., 2023), and beyond. Economic tools are likewise being applied to generative AI models (Chen et al., 2023; Kim et al., 2024).

[3]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

Fudenberg et al. (2022) and Fudenberg et al. (2025) respectively develop the measures of completeness and restrictiveness, which we adopt here to evaluate a model's prediction accuracy and flexibility.[4] 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 on certainty equivalents. Building on this analysis, Fudenberg et al. (2025) show that CPT achieves much higher completeness than a two-parameter specification generated by DA, proposed by Gul (1991), but CPT is also substantially less restrictive. Similarly, Fudenberg and Puri (2021) and Fudenberg and Puri (2022) evaluate the completeness of multiple EUT and non-EUT specifications with and without simplicity preferences (Puri, 2025).

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. Previous studies primarily evaluate prediction accuracy and flexibility based on a small number of individual decisions within relatively constrained choice scenarios. In contrast, we present subjects with choices subject to a budget constraint, which provide richer information about preferences than typical discrete-choice settings. Moreover, we elicit many such choices, yielding a substantially larger dataset. This enables analysis at the level of the individual subject, without pooling data or assuming homogeneity across subjects. Most importantly, because choices are drawn from standard budget lines, we can apply classical revealed preference analysis to assess whether 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.

trigger strategies perform well relative to ML models in predicting cooperation rates in repeated games. Other papers, such as Hsieh et al. (2025) and Peterson et al. (2021), conduct the same type of predictive exercise between economic models of choice under risk using neural networks as Bernoulli utility functions.

[4]Rather than evaluating predictive performance, Ludwig and Mullainathan (2024) and Mullainathan and Rambachan (2024) employ generative adversarial methods that generate synthetic observations to optimize an objective under realism constraints. While we do not adopt these methods in the present work, they represent promising directions for future research.

3 Framework for analysis

Next, we outline the framework for our analysis. We begin with an overview of the experimental design, procedures, and data. We then discuss completeness and restrictiveness, our primary measures of interest. Following this, we present the revealed preference tests of GARP and FOSD. Finally, we provide details on the economic and ML models we evaluate.

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 line 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. Also let $x = x_1/(x_1 + x_2)$ denote the relative demand, or token share, of the security that pays off in state 1. The individual-level dataset generated by a subject's choices from linear budget lines is then given by $\{(\mathcal{B}^i, x^i)\}_{i=1}^{50}$, where $\mathcal{B}^i$ denotes the i-th observation of the budget line and x^i denotes the corresponding token share. Let $\boldsymbol{\mathcal{B}}$ denote the set of budget lines.[5]

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.

To choose an allocation, subjects used the mouse or the arrows on the keyboard

[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 budget line in decision round $i = 1, ...50$. Without loss of generality, the income m is normalized to 1.

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. The experimental instructions are available in the Online Appendix.

Our dataset consists of 956 subjects, including participants from the symmetric (equal probability) experiment of Choi et al. (2007a), identical experiments conducted by Zame et al. (2026) and Cappelen et al. (2023), as well as other previously collected data. The budget line graphical interface was introduced by Choi et al. (2007b), applied by Choi et al. (2007a) with student subjects, and later by Choi et al. (2014) with a nationally representative sample. These datasets have subsequently been analyzed in numerous papers, including Halevy et al. (2018), Polisson et al. (2020), de Clippel and Rozen (2023), and Echenique et al. (2023), among others.[6] In addition, following Fisman et al. (2007), a series of papers have employed a similar methodology to study social preferences using different subject pools. As such, the dataset in this paper forms part of a substantial and well-known body of data familiar to experimentalists.

3.2 Measures

Following the terminology and notation of Fudenberg et al. (2025), 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

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

[6] We do not use the Choi et al. (2014) dataset here, because each subject made only 25 decisions. Although this exceeds what is typical in experiments, and revealed preference analysis shows that the variation in budget lines (prices and incomes) is sufficient to permit a rigorous test of consistency with 25 decisions, we rather focus on datasets with 50 decisions for ML applications. Ellis (2025) uses the dataset of Choi et al. (2014) to investigate the predictive value of auxiliary covariates.

where P denotes the joint distribution of $(\mathcal{B}, x)$.[7] 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. (2025) 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.

- The restrictiveness of a model discerns 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.

The completeness and restrictiveness of *nested* models, such as EUT and RDU, can be easily compared—the completeness/restrictiveness of a nested model (EUT) can be no higher/lower than its nesting model (RDU). In practice, however, 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

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

[7] Note that the marginal over budget lines 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 line $\mathcal{B}$.

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 lines at most once, it is possible to construct a function from budget lines 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 the model's 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. (2025) 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 Testing GARP and FOSD

The most basic question to ask about choice data is whether it is consistent with individual utility maximization. If budget lines 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 lines 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 satisfy GARP.[8]

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 comply with GARP by calculating Afriat's (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 lines 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 problem at hand. Given that 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.[9]

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.[10] Clearly, always allocating all tokens to one of

[8] 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 provide a stringent test of utility maximization.

[9] 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.

[10] 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 accounts generates severe violations of GARP in the combined dataset, 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.

3.4 Economic models

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 contains EUT as a special case 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 $u(\cdot)$ 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 curve, 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).

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. Hence, utility is a weighted

average, with weights β_L and β_H, of the expected utility when the payoffs are x_L and x_H, respectively. We note that with two equally likely states the RDU functional form above can arise from different classes of non-EUT preferences such as DA, proposed by Gul (1991), where the utility function takes the form

$$\min\left\{\alpha u\left(x_1\right)+u\left(x_2\right), u\left(x_1\right)+\alpha u\left(x_2\right)\right\},$$

where $\alpha \geq 1$ is a parameter measuring loss/disappointment aversion (and the safe allocation $x_1 = x_2$ is taken to be the reference point), which is exactly the RDU functional form with an appropriate change of variables.

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.

3.5 Machine learning models

Our analysis spans eight models, grouped into three principal families of ML techniques: regularized regressions (OLS, Lasso, Ridge Regression), tree-based methods (Mean tree, Linear model tree, Support Vector Regression tree, Random Forest), and neural networks. Each family is widely applied in the ML literature and is increasingly prominent in economics research. We adopt multiple models, reflecting the absence of a clearly recognized 'winning' method.[11] 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 briefly describe our application of the

[11] 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 ..." More recent work by McElfresh et al. (2024) find this remains to be the case for tabular data, which is the format we consider here.

models below and refer readers to the Online Appendix for further details.[12]

Regularized regressions Regularized regression, in its simplest form, assumes a linear relationship between outcomes and covariates, whose coefficient is estimated using the objective function of Ordinary Least Squares (OLS) 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. In addition to OLS, which we include as an unregularized baseline, we consider two popular models of regularized regression that add the norm of the coefficient vector as the penalty, which differ in which norm is implemented as the penalty. First, we consider Lasso (Tibshirani, 1996), which penalizes using the L_1 norm. Second, we consider Ridge Regression (Hoerl and Kennard, 1970), which penalizes using the L_2 norm. The norm is multiplied by a parameter λ, which controls the degree to which coefficient magnitudes influence the objective function.

Tree-based Unlike the global linear relationship assumed in regression models, tree-based models partition the set of budget lines $\boldsymbol{B}$ into subsets (based on the prices and the endowment) and estimate a submodel on each of the subsets. The resulting tree-based model is thus a piecewise function with each partition having a separately applied submodel. Partitioning is done recursively. That is, given some subset of budget lines, the model considers a further binary partition that minimizes the size-weighted error of both partitions.[13]

The standard decision tree submodel, denoted Mean, takes the sample mean token share x of each subset. We use Mean as well as three extensions. The first two extensions, known more broadly as model trees (Quinlan et al., 1992), change the estimated submodel from a sample mean to a Linear Regression, and Support Vector Regression (SVR) with a normal radial basis function. The last tree-based model, the

[12]For readers less familiar with these methods, Hastie et al. (2009) and Daumé (2017) provide accessible textbook introductions. It should be taken into account, however, that this is a fast-moving field.

[13]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 by setting a minimum number of observations per partition and limiting the "depth", or number of partitions away from $\boldsymbol{B}$, of a tree. Exact details can be found in the Online Appendix.

Random Forest (RF), averages the decision rules of multiple standard decision trees. Each tree is given a bootstrapped dataset, and is generally seen as an improvement over singular decision trees (Breiman, 2001).

Neural networks Neural networks, specifically multilayer perceptrons, transform budget lines into relative demand by nonlinear regression, whose functional form assumes a series of nested transformations. Each transformation consists of an affine transformation and nonlinear transformation, whose output is used as the input to the following transformation. 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. As described in detail in the Online Appendix, we use the layer count, layer dimension, and $\sigma^{(i)}$ values from Hsieh et al. (2025).

4 Results

We first present the individual-level analysis. We then compare these results—which constitute the main contribution of the paper—to aggregate-level findings common in the literature, as well as to clustering-level results where completeness is calculated for each cluster–model pair. Finally, we turn to simulated subjects, who are able to make many more choices than human subjects, to assess the robustness of our analysis.

4.1 Individual-level analysis

Table 1 provides a population-level summary of our results, complementing the information provided in Figure 1. The left column of Table 1 reports the average completeness of each model, along with the 95% confidence interval for that average. The next column presents the win rate—that is, the fraction of subjects for whom EUT is more complete—against each model. The next two blocks of four columns present the win rate of EUT against each model, along with the absolute completeness differences by quartiles of e^{**}, the consistency score with GARP and FOSD. The rightmost column reports the restrictiveness of each model.

The upper panel of Table 1 presents the results for EUT, RDU, and the subject-

specific best-performing ML model. The lower panel reports the results for the three principal families of ML models: regularized regressions, tree-based approaches, and neural networks. For regularized regressions and tree-based models, we report restrictiveness as the weighted average of the most complete model in each class for each subject. The Online Appendix provides the results for the three regularized regression models and the four tree-based models.

In comparing EUT with RDU, which nests it, we observe that the average completeness of EUT is not lower than that of RDU (89.3% versus 89.2%). Moreover, EUT is more restrictive (18.6% versus 16.6%), consistent with the fact that RDU embeds EUT as a parsimonious special case where $\beta_L = \beta_H$. The overall win rate of EUT relative to RDU is 60.6%. While consistently above 50%, this win rate declines from 68.3% in the first quartile of e^{**} to 53.6% in the fourth. The absolute completeness differences between EUT and RDU are small, except in the first quartile of e^{**} scores, which corresponds to the least consistent subjects with GARP and FOSD. While EUT's advantage in completeness is modest for most subjects, we nonetheless view this as a meaningful 'victory' given the model's parsimony. We therefore focus below on the comparison of EUT to ML models.

Three main insights emerge from Table 1 regarding the completeness and restrictiveness of EUT compared with ML models:

- First, the completeness of EUT is comparable to that of tree-based models (89.3% versus 89.1%), but it is significantly higher than that of regularized regression models and neural networks (79.5% and 71.6%, respectively). Furthermore, EUT's completeness win rate rises from 69.7% against tree-based models to 88.5% against regularized regression models and 94.2% against neural networks.

- Second, the win rate of EUT almost always increases across e^{**} consistency quartiles against all three families of ML models, as does its relative improvement in terms of absolute completeness differences over regularized regression models and neural networks. Importantly, the predictive accuracy of EUT improves relative to ML models when individual choices more closely satisfy GARP and FOSD—the axioms on which both EUT and non-EUT models of choice under risk are based.

- Third, while EUT does not achieve a large improvement in completeness com-

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

Model classes	Average completeness	EUT win rate	EUT's win rate against models by e^{**} quartiles				Absolute completeness difference between EUT and models by e^{**} quartiles				Restrictiveness
			1st	2nd	3rd	4th	1st	2nd	3rd	4th	
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
Best ML	89.6 [88.9, 90.3]	65.1	52.9	67.8	66.2	73.4	-2.8	0.6	0.4	0.6	11.7
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

The left column reports the average completeness of each model, along with the 95% confidence interval calculated via nonparametric bootstrap: for each model, we sample the observations with replacement 10,000 times, compute the mean on each resample, and take bias-corrected 95% interval. The next column reports the win rate of EUT (the fraction of subjects for whom EUT is more complete) against each model. The following two blocks of four columns report the win rate of EUT against each model and its absolute completeness difference by quartiles of e^{**}, the consistency score with GARP and FOSD. The right column reports the restrictiveness of each model. For regularized regressions and tree-based models, restrictiveness is reported as weighted averages of the most complete model in the class for each subject. Results for the three regularized regression models and the four tree-based models are provided in the Online Appendix.

pared with tree-based models, it is substantially more restrictive (18.6% versus 10.6%). Moreover, the restrictiveness of EUT is comparable to that of regularized regression models and neural networks (20.7% and 14.4%, respectively), but these ML models are significantly less complete than EUT.

The final conclusion from Table 1 is that, even when the best ML model is selected for each subject to compete against EUT, the preceding results remain largely robust under this demanding test.[14] Although EUT is marginally less complete than the most complete ML model (89.3% versus 89.6%), its win rate of 65.1% indicates that it delivers higher completeness for the majority of subjects. Moreover, EUT is substantially more restrictive (18.6% versus 11.7%), and its relative performance improves among subjects whose choices are more consistent with GARP and FOSD, as evidenced by the results across e^{**} quartiles. The anatomy of the economic models' success—that is, the explanation for why they outperform ML models—lies in the consistency of individual choices with the fundamental axioms of GARP and FOSD, which underlie both EUT and non-EUT models.

To illustrate this, we focus on subjects whose behavior aligns with several prototypical notions of risk aversion. It should be noted, however, that for the majority of subjects the data exhibit considerably less regularity. Even so, the choices observed across the full sample reveal pronounced consistency within subjects and substantial heterogeneity across subjects. Table 2 presents the average completeness scores of EUT, RDU, and the most complete ML model for 55 subjects (5.8%), 42 subjects (4.4%), and 30 subjects (3.1%) whose choices are nearly consistent with maximizing prototypical EUT preferences—namely infinite risk aversion, risk neutrality, and a logarithmic von Neumann-Morgenstern utility function. These subjects typically selected the safe allocation on the 45-degree line ($x_1 = x_2$), the highest intercept with the axes ($x_s = \frac{m}{p_s}$ if $p_s \leq p_{s'}$), and the midpoint of the budget line segment ($x_s = \frac{m}{2p_s}$), respectively. In addition, we report results for 36 subjects (3.8%) whose choices align with loss/disappointment aversion (and risk neutrality). These subjects generally chose the highest intercept with the axes on steep or flat budget lines, and the safe allocation on the 45-degree line for intermediate budget lines corresponding to $\ln(p_1/p_2)$ around zero. Such behavior departs from EUT, instead exhibiting loss/disappointment aversion in which the safe allocation is the reference point. The

[14] The best ML model is regularized regressions, tree-based approaches, and neural networks for 17.9%, 77.1%, and 5.0% of subjects, respectively

classification of subjects allows for a narrow confidence interval that accounts for minor inaccuracies from subjects' mouse handling, and the results remain robust under alternative confidence interval specifications.

Table 2: The completeness of EUT, RDU, and the most complete ML model for subjects with prototypical preferences

	Average completeness			EUT win rate (%)		
Demand pattern	EUT	RDU	Best ML	vs. RDU	vs. Best ML	Obs
Infinite risk aversion	99.0	98.7	99.1	67.3	54.6	55 (5.8%)
Risk neutrality	99.6	99.5	97.5	69.1	78.6	42 (4.4%)
Log utility	98.4	98.4	97.7	66.7	83.3	30 (3.1%)
Loss/disappt. aversion	94.9	95.8	95.9	30.6	47.2	36 (3.8%)

Each row corresponds to prototypical preferences. The first three - infinite risk aversion, risk neutrality, and log utility - are consistent with EUT, whereas the last - loss/disappointment aversion - is only consistent with RDU. The first block of three columns reports the average completeness of EUT, RDU, and the most complete ML model, respectively. The next block of two columns reports the win rate of EUT against RDU and the most complete ML model, respectively. The final column reports the number and fraction of subjects whose choices are consistent with each prototypical preferences type. The results allow for a narrow confidence interval to accommodate small mistakes resulting from slight imprecision in subjects' handling of the mouse, and they are robust to different confidence interval specifications.

The first three rows of Table 2 reveal a consistent pattern: when individual choices are consistent with EUT, the added flexibility of ML models does not enhance predictive accuracy and, in many cases, yields slightly worse performance—even with extensive individual-level data, as in our experiment. In these special cases—consistent with maximizing a utility function of the expected utility form—the same applies to the more flexible economic model, RDU. Because additional flexibility cannot improve completeness for these subjects, EUT and RDU display nearly equivalent performance. In contrast, when behavior aligns with RDU but not with EUT, as in the final row, both RDU and the best ML model attain substantially higher completeness and outperform EUT for the vast majority of subjects.

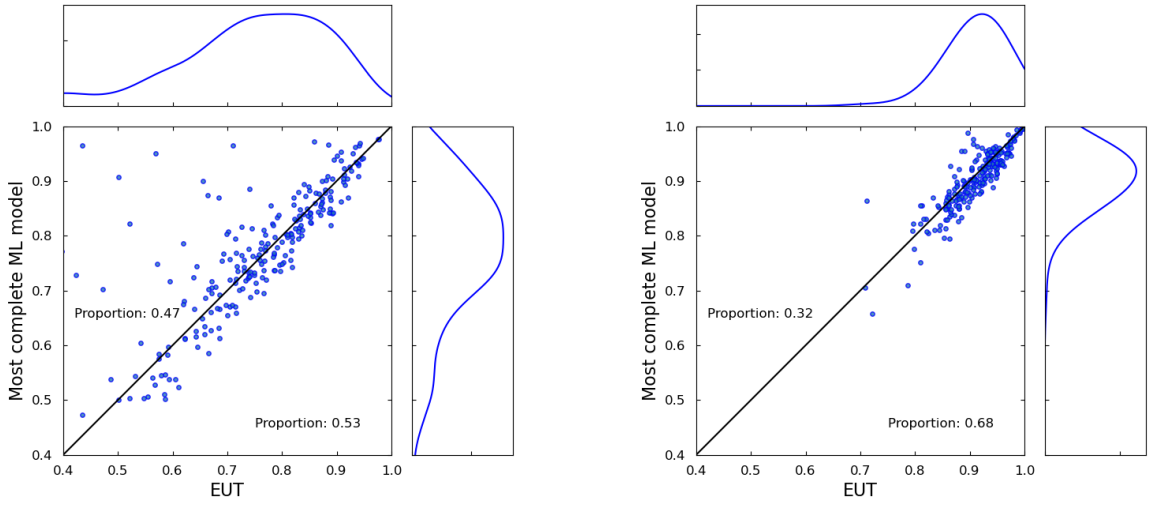
We further identify a small subset of subjects for whom the best ML model significantly outperforms the economic models. The choices of these subjects are inconsistent with the fundamental axioms of ordering and/or monotonicity as measured by e^{**}—but in a systematic way. A straightforward example is the two subjects

(0.2%) who always chose one of the intercepts, either on the x-axis or the y-axis. Their behavior satisfies GARP ($e^* = 1$) but severely violates FOSD (low e^{**} scores), as they selected the lower intercept in half of the decision rounds, on average. The completeness scores for these two subjects are 24.6% and 24.4% for EUT, 25.3% and 21.0% for RDU, compared with above 99.9% for the best ML model. It is worth noting that among our nearly thousand subjects, no subject always chose the lower intercept violating both GARP and FOSD, which further indicates subjects' comprehension of the experiment.

To broaden the discussion, Figure 2 presents a detailed comparison of completeness between EUT and the most complete ML model across the entire sample. The four panels correspond to quartiles of e^{**}, reflecting the consistency of individual-level data with GARP and FOSD. Within each panel, the horizontal axis depicts the completeness of EUT and the vertical axis depicts the completeness of the best ML model. Marginal kernel density estimates of completeness scores, approximated using a Gaussian kernel, are displayed along both axes to highlight the distributional patterns. We also show the proportion of subjects that lie above and below the diagonal, which correspond to the proportion of subjects for which ML or EUT have higher completeness, respectively.

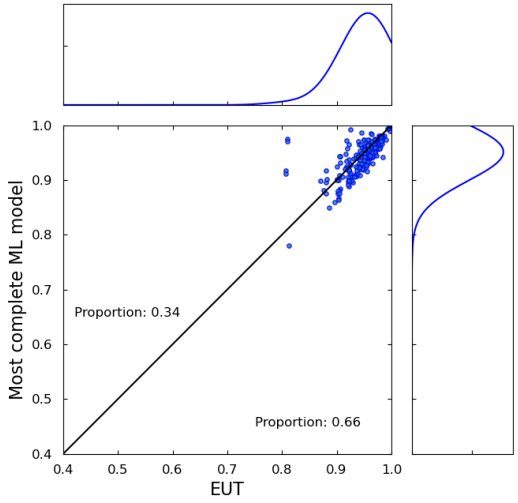
We first note that extreme differences in completeness are rare, as evidenced by the absence of observations in the upper-left and lower-right corners of each panel. The notable exception occurs in the lowest e^{**} consistency quartile (Panel a), where observations above the diagonal show the best ML model substantially outperforming EUT for subjects whose choices are inconsistent with GARP and/or FOSD, in line with the earlier example. Beyond this, we observe a monotonic shift toward the upper-right corner across e^{**} quartiles, reflecting greater completeness of both EUT and ML models as individual choices become more consistent. The proportion of observations below the diagonal—subjects for whom EUT is the most complete model—weakly increases across e^{**} quartiles, and EUT exhibits higher completeness distributions in all panels. Finally, we note that among the 334 subjects for whom the most complete ML model surpasses EUT, 245 (73.3%) have EUT as the second-most complete model, outperforming the other ML models.

Figure 2: Individual-level scatter plots of the completeness of EUT versus the most complete ML model by e^{**} quartile

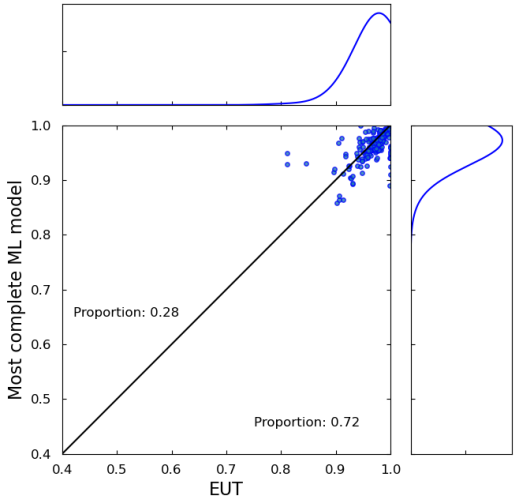


(a) e^{**} quartile 1

(b) e^{**} quartile 2



(c) e^{**} quartile 3



(d) e^{**} quartile 4

The individual-level completeness scores of EUT (horizontal axis) versus the most complete ML model (vertical axis), by quartile of e^{**} - the consistency score with GARP and FOSD. Panel (a) plots the subjects in the lowest quartile, and Panel (d) those in the highest. The quartile ranges are $[0, 0.83)$, $[0.83, 0.95)$, $[0.95, 0.99)$, and $[0.99, 1]$. Each axis also includes a marginal kernel density estimate of completeness scores, approximated using a Gaussian kernel.

4.2 Aggregate and clustering-level comparisons

Recognizing that multiple levels of heterogeneity may arise between the individual-level analysis employed here and the aggregate-level common in the literature, we next compare our results across individual-level analysis, aggregate-level analysis, and clustering-level analysis with two clustering specifications. For each subject, we compute the average fraction of tokens allocated to the cheaper of the two assets, $x_L/(x_L + x_H)$. This nonparametric statistic of risk-taking is then used to generate either five or ten clusters via k-means clustering, after which completeness is estimated for each cluster–model pair.[15]

Two methodological points are worth noting in estimating completeness at the pooled level. First, subjects are stratified to ensure that their data contribute equally to each fold of cross-validation. Second, because completeness is a nonlinear operator, averaging completeness scores across subjects differ from calculating completeness based on the average model error across subjects. To ensure comparability with the individual-level analysis, we adopt the former approach.

Table 3 presents average model completeness across different levels of analysis. The first column reproduces the individual-level completeness reported in Table 1. The second and third columns report average completeness under clustering specifications of ten and five clusters, respectively. The fourth column reports completeness when all subjects are pooled together. Confidence intervals are derived from bootstrap estimates of the mean.

The results in Table 3 highlight two principal trends. First, at the individual level, EUT is comparable in performance to the most complete ML model (89.3% versus 89.9%). Once data are clustered, however, ML models become relatively more complete than EUT, and at the aggregate level the most complete ML model significantly outperforms EUT (83.1% versus 80.3%). Second, completeness declines across all models as the analysis level coarsens from left to right (with the exception of neural networks, which decline except for the transition from individual-level analysis to 10-cluster analysis). Although increasing the number of clusters naturally

[15]For both EUT and RDU, parameters are estimated from the pooled data of subjects within each cluster, and these estimates are then used to predict demand out of sample. To accommodate the larger dataset size, the parameter search space for tree-based models is expanded accordingly (see Online Appendix for details). Clearly, both individual-level and aggregate analyses admit a clustering interpretation: in the former, each cluster is a singleton, whereas in the latter the entire sample may be regarded as a single cluster.

Table 3: The completeness of EUT, RDU, and ML models at the individual, clustered, and aggregate levels

	Completeness			
Models	Individual	Ten clusters	Five clusters	Aggregate
EUT	89.3	88.4	87.7	80.3
	[88.4, 90.0]	[87.5, 89.1]	[86.8, 88.4]	[79.2, 81.3]
RDU	89.2	88.4	87.8	80.9
	[88.3, 90.0]	[87.5, 89.1]	[87.0, 88.6]	[79.9, 81.8]
Best ML	89.6	89.0	89.0	83.1
	[88.9, 90.3]	[87.9, 89.9]	[87.9, 90.0]	[81.5, 84.4]
Regularized regressions	79.5	68.6	68.1	61.1
	[77.7, 80.6]	[66.1, 71.0]	[65.6, 70.6]	[58.0, 64.0]
Tree-based models	89.1	88.4	88.1	75.1
	[88.4, 89.9]	[87.3, 89.3]	[86.8, 89.1]	[73.4, 76.6]
Neural networks	71.6	87.0	86.9	73.0
	[68.8, 73.6]	[85.9, 88.0]	[85.7, 88.0]	[71.1, 74.7]

Each column reports the average completeness of each model, along with the 95% confidence interval. The first column replicates the first column of Table 1, reporting the average individual-level completeness score. The second and third columns report the average completeness when subjects are clustered using k-means clustering on the average fraction of tokens allocated to the cheaper asset, with k set to 10 and 5, respectively. The final column reports aggregate-level completeness. We compute mean completeness intervals via nonparametric bootstrap: for each model-cluster level pair, we sample the observations with replacement 10,000 times, calculate the mean on each resample, and take the bias-corrected 95% interval as the confidence interval.

reduces the data available per cluster, the gains in predictive precision dominate. Interestingly, although the best ML model surpasses EUT at both the clustered and aggregate levels, individual families outperform EUT in only one instance: tree-based models at the five cluster level. Even in this case, the improvement is marginal (87.7% versus 88.1%), as illustrated in the bottom panel of Table 3. Finally, RDU is somewhat more complete than EUT at the aggregate level (80.9% versus 80.3%)—consistent with prior literature employing aggregate-level analysis—but it does not outperform EUT at the individual or clustered levels.

Clearly, even a high degree of consistency in individual-level decisions with GARP and FOSD does not imply consistency with these fundamental economic axioms in

pooled data. The considerable heterogeneity in subjects' behavior means that while choices may be individually consistent, they are mutually inconsistent—even when subjects are clustered by levels of risk attitudes. The analysis shows that the higher completeness of ML models relative to economic models reported in prior studies—and replicated when pooling our own data—stems from the ML models' ability to capture systematically inconsistent patterns arising from individual heterogeneity in aggregate-level data.

4.3 Analysis with simulated subjects

While the number of individual decisions in our experiments exceeds that typically reported in the experimental literature—yielding a rich dataset of choices across diverse budget lines—it remains important to evaluate the completeness of economic models relative to ML models in scenarios where the number of decisions is substantially greater than what a human subject could reasonably undertake within a single experimental session.

To this end, we generate a random sample of simulated subjects who implement either EUT or RDU (with distortion $w\left(1/2\right)=2/3$) 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). The probability of a specific allocation $\mathbf{x}$ being chosen from a budget line with prices $\mathbf{p}$ is thus:

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

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

We constructed samples of simulated subjects with CRRA risk-aversion parameter $\rho=1/2$, consistent with the range estimated for our human subjects, and four precision levels: $\gamma=0,0.25,1,10$. Each simulated subject makes 1,000 choices from randomly generated budget lines, whereas each human subject made 50 choices. Figure 3 presents average completeness for EUT, RDU, and the most complete ML model across training-set checkpoints of size 25, 50, 100, 250, 500, 750, and 900. The

left panel corresponds to simulated subjects maximizing an EUT utility function, and the right panel to those maximizing an RDU utility function. Rows indicate different levels of logistic error, with decreasing precision parameter γ; the bottom row, where $\gamma = 0$, represents uniform choice across budget lines. Each bar in Figure 3 reports the average completeness across 1,000 simulated subjects. Confidence intervals, as calculated throughout the paper, are 95% bootstrapped confidence intervals of the mean, but we omit them because they are too narrow to be visually discernible.

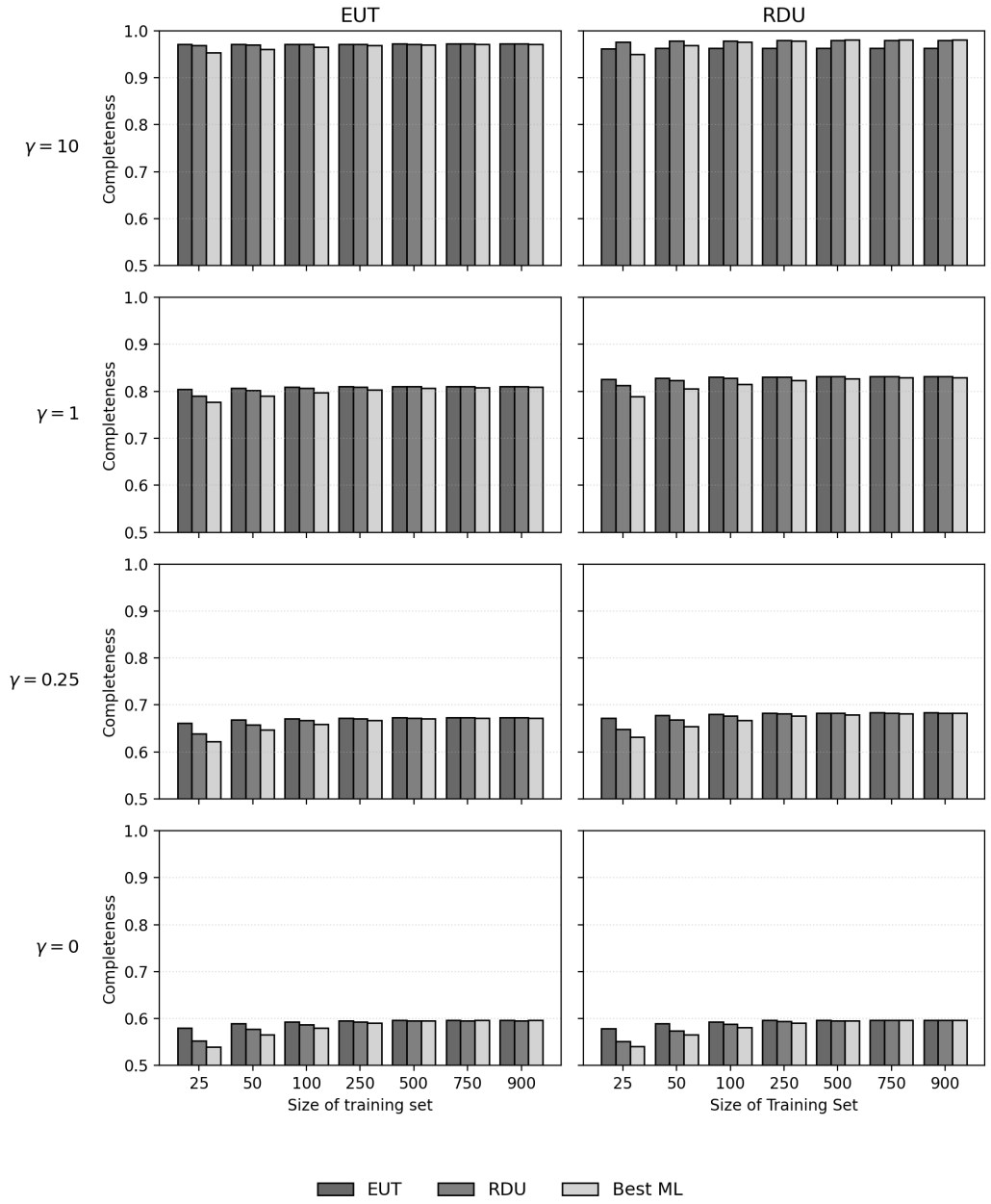
Figure 3 yields several key insights. First, the completeness results for our actual subjects align with the higher-precision simulations $\gamma = 1$ and $\gamma = 10$ using 50 observations. Second, under an EUT data-generating process (left panels), EUT is at least as complete as RDU and the most complete ML model across all sample sizes and noise levels, with its relative advantage increasing as sample size decreases and noise rises. Third, under an RDU data-generating process (right panels), RDU is the most complete model in low-noise settings, as expected. Yet EUT surpasses RDU when sample sizes are small and noise is high. Importantly, EUT is significantly less complete than RDU only under low-noise conditions, and it generally remains more complete than the best-performing ML model on RDU-generated data—except in the case of large sample sizes with low noise.

To conclude, Figure 3 shows that the relative completeness of EUT compared to ML models is preserved even when subjects are simulated with substantially larger individual-level datasets than those used in our experiments. Crucially, this advantage does not diminish with larger datasets; instead, the completeness of EUT proves robust to increases in the number of individual decisions. Indeed, the average completeness scores of EUT, RDU, and the most complete ML model across all noise levels nearly converge at roughly 50 choices. We also obtain similar results when introducing noise by randomly and uniformly reallocating varying fractions of choices across each budget. To economize on space, we omit these results but they underscore the robustness of the simulated-subject analysis in this part of the paper.

5 Conclusion

We compare the predictive performance of standard economic models of choice under risk and machine learning (ML) models, using data from nearly 1,000 subjects who made choices through graphical representations of budget lines over bundles of

Figure 3: The completeness of EUT, RDU, and the most complete ML model for simulated subjects



Data are generated for simulated subjects making decisions on budget lines as human subjects did. We report the completeness of EUT, RDU, and the most complete ML model when trained on 25, 50, 100, 250, 500, 750, and 900 decisions. Each bar indicates the average completeness of 1,000 simulations. The 95% confidence intervals, calculated via nonparametric bootstrap, are very narrow and therefore omitted. The left (resp. right) panels correspond to simulated subjects maximizing an EUT (resp. RDU) utility function with decision noise. Each row represents a different level of logistic error, with decreasing precision parameter γ (where the bottom panel, with $\gamma = 0$, corresponds to choosing uniformly on each budget line).

state-contingent commodities, enabling the collection of rich individual-level data. Our main finding is that the economic models outperform all ML models, with a wider margin as individual choices become more consistent with GARP and/or FOSD. We view this as a 'victory' for the economic models, particularly EUT, as it is nested within RDU and thus more restrictive. The rare cases in which ML significantly outperforms the economic models involve subjects whose choices violate GARP and/or FOSD, but do so systematically.

In contrast to our individual-level results, the best ML model outperforms the economic models when data are pooled across all subjects, consistent with prior literature on aggregate completeness. This occurs because individual consistency with GARP and FOSD does not yield aggregate consistency due to heterogeneity. Clustering subjects and calculating cluster–model completeness scores shows that, as analysis levels become coarser, ML models gain relative completeness over the economic models. In addition, while RDU is more complete than EUT at the aggregate level, it does not outperform EUT at the clustered levels. Our final analysis, based on simulated subjects who make far more choices than human subjects can, shows that our main result—that the economic models outperform all ML models—is robust even with much larger individual-level datasets.

We view this as the beginning of a broader agenda that builds on the experimental methodology and analytical techniques developed in this paper. Examining behavior in more complex settings will naturally require additional experimental data as well as new theoretical and methodological tools. Below, we briefly discuss current extensions of our work:

- Ellis et al. (2024a) examine richer data from three-dimensional budget lines, where the axioms underlying economic models impose more — and more stringent — restrictions on observed behavior. We draw on data for choice under risk from Dembo et al. (2026) and under ambiguity from Ahn et al. (2014). Consistent with the results reported here, we find that EUT under risk and Subjective Expected Utility (SEU) under ambiguity are at least as complete as non-EUT and non-SEU models and outperform ML models. Relative to the two-dimensional case, the completeness of economic models is only slightly reduced, while their restrictiveness increases substantially. Overall, economic models remain more complete than ML models and are considerably more restrictive.

- Ellis et al. (2024b) examine the transferability of models from two - to three-dimensional budget lines. Using within-subject data, we test whether models estimated in a simpler two-state risky environment can predict choices in a more complex three-state setting.[16] We find substantial transferability for most subjects: EUT retains over 90% of its within-domain predictive accuracy when generalized, outperforming both non-EUT and ML models. These findings highlight the robustness of parsimonious economic models, particularly EUT, in extrapolating reliably from simple to more complex domains.

- Ellis (2025) investigates the predictive value of auxiliary covariates in our setting, using the Choi et al. (2014) dataset from the general population. The analysis simulates varying levels of data availability by selectively removing demographic covariates, subject identifiers, or both. EUT serves as the benchmark, with its out-of-sample predictive performance compared to ML models. The main finding is that identifying information is more valuable than demographic data, though both significantly improve prediction relative to choice data alone. EUT remains competitive with ML models, particularly outperforming them for subjects whose choices are consistent with GARP and FOSD.

[16] Andrews et al. (2024) develop measures of cross-domain transfer and find, consistent with the findings of Ellis et al. (2024b), that economic models of choice under risk outperform ML models when predicting certainty equivalents of binary lotteries, primarily due to their ability to extrapolate to different payoff values.

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.

Ahn, D., S. Choi, D. Gale, and S. Kariv (2014): "Estimating ambiguity aversion in a portfolio choice experiment," *Quantitative Economics*, 5, 195–223.

Almog, D., R. Gauriot, L. Page, and D. Martin (2024): "AI Oversight and Human Mistakes: Evidence from Centre Court," in *Proceedings of the 25th ACM Conference on Economics and Computation*, 103–105.

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): "Machine Learning Methods Economists Should Know About," *Annual Review of Economics*, 11, 685–725.

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

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

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

Brynjolfsson, E., D. Li, and L. Raymond (2025): "Generative AI at Work," *The Quarterly Journal of Economics*, 140, 889–942.

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 (2025): "The Next Generation of Experimental Research with LLMs," *Nature Human Behaviour*, 1–3.

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*, Alanna Maldonado.

de Clippel, G. and K. Rozen (2023): "Relaxed optimization: How close is a consumer to satisfying first-order conditions?" *Review of Economics and Statistics*, 105, 883–898.

Dembo, A., S. Kariv, M. Polisson, and J. K.-H. Quah (2026): "Ever Since Allais," *Journal of Political Economy*, Forthcoming.

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

Echenique, F., T. Imai, and K. Saito (2023): "Approximate expected utility rationalization," *Journal of the European Economic Association*, jvad028.

Ellis, K. (2025): "The Value of "Who" and "What" When Predicting Choice Under Risk," *Working Paper*.

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

Ellis, K., S. Kariv, and E. Y. Ozbay (2024b): "Scaling Up: Individual-Level Transfer Performance of Models," *Working Paper*.

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

Fisman, R., S. Kariv, and D. Markovits (2007): "Individual preferences for giving," *American Economic Review*, 97, 1858–1876.

Fudenberg, D., W. Gao, and A. Liang (2025): "How Flexible Is That Functional Form? Quantifying the Restrictiveness of Theories," *The Review of Economics and Statistics*, 1–16.

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 (2021): "Evaluating and Extending Theories of Choice Under Risk," *Working Paper*.

——— (2022): "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.

Halevy, Y., D. Persitz, and L. Zrill (2018): "Parametric recoverability of preferences," *Journal of Political Economy*, 126, 1558–1593.

Hastie, T., R. Tibshirani, 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 (2025): "Logit Neural-Network Utility," *Journal of Economic Behavior and Organization*, 236, 107054.

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*.

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.

McElfresh, D., S. Khandagale, J. Valverde, V. Prasad C, G. Ramakrishnan, M. Goldblum, and C. White (2024): "When do neural nets outperform boosted trees on tabular data?" *Advances in Neural Information Processing Systems*, 36.

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

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. (2025): "Simplicity and risk," *The Journal of Finance*, 80, 1029–1080.

Quiggin, J. (1982): "A Theory of Anticipated Utility," *Journal of Economic Behavior and 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.

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.

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 (2026): "Linking Social and Personal Preferences: Theory and Experiment," *Journal of Political Economy*, Forthcoming.