What can the demand analyst learn from machine learning?*

Keaton Ellis, Shachar Kariv and Erkut Ozbay†

November 28, 2022

Abstract

We compare the predictive performance of a standard economic model to a variety of machine learning models by presenting nearly 1,000 subjects with a consumer decision problem – the selection of a bundle of contingent commodities from a budget set. Our dataset allows us to compare predictions at the individual level and relate them to the consistency of individual decisions with revealed preference axioms. Using dual measures of completeness and restrictiveness from Fudenberg and Liang (2019), we show that the economic model outperforms all machine learning models, with a wider margin as choices align more with an underlying preference ordering.

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

1 Introduction

The economic theory of consumer behavior assumes the decision-maker has consistent (complete and transitive) preferences over all possible alternatives and chooses the most preferred alternative from the feasible set. Applied demand analysis, therefore, addresses four types of questions (Varian (1982), Varian (1983)): (*i*) *Consistency*. Is behavior consistent with a model of utility maximization? (*ii*) *Structure*. Does the rationalizing utility function have some special structural properties? (*iii*) *Recoverability*. Can the underlying preferences be recovered from observed choices? (*iv*) *Extrapolation*. How can we forecast behavior in other circumstances?

In the economic approach, the demand analyst, therefore, tests whether behavior can be rationalized by some preference ordering (or posit a utility function with some special structure), derive the associated demand function, and fit it to data using some econometric technique. There is a

*We are grateful to Annie Liang and Sendhil Mullainathan for encouragement and helpful discussions, and to a number of seminar audiences for helpful suggestions.

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

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 forecast behavior. By now such analysis is 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 models are built *solely* for the purpose of extrapolation by seeking functions that minimize out-of-sample prediction error. As pointed by Mullainathan and Spiess (2017), among others, machine learning (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 demand behavior. To fix ideas, consider a sequence of standard consumer decision problems – the selection of a bundle of commodities from a standard budget set. 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 consumer'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 $(\mathbf{p}^i, \mathbf{x}^i)$?

The key dual concepts in this regard are *completeness* and *restrictiveness* (Fudenberg et al. (2021)). 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 this paper, we compare the completeness and restrictiveness of economic models to that of a variety of ML models using data from an economically important experimental setting that can be interpreted as a portfolio choice problem – the selection of a bundle of contingent commodities from a standard budget set. These decision problems are presented using the graphical experimental interface of Choi et al. (2007b). 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]

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

[1]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). We, therefore, argue that it is clearly advantageous to estimate individual-level parameters and then generate individual-level distributions rather than to pool data and estimate type-level parameters.

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 a thousand subjects. For each subject, we have 50 observations $\{(\mathbf{p}^i, \mathbf{x}^i)\}_{i=1}^{50}$ over a wide range of budget sets.[2]

For each subject, we calculate the completeness and restrictiveness of the Rank Dependent Utility (RDU) model of Quiggin (1992, 1993), which encompasses a number of prominent economic models of choice under risk. von Neumann and Morgenstern's (1947) Expected Utility Theory (EUT) is a special case of this model. We then compare, *subject-by-subject*, the completeness of this economic model and the *most* complete prediction model among *eight* ML models across *three* main families of ML models – regularized regressions, tree-based, and neural networks. For each subject, we also 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). RDU weakens the independence axiom but embodies ordering (completeness and transitivity) and monotonicity with respect to FOSD, as are almost all models that generalize EUT.

Figure 1 depicts our main result. The horizontal axis presents quartiles of consistency scores with GARP and FOSD. The vertical axis indicates the fraction of subjects for whom RDU is *more* complete than the *most* complete ML model within each class – regularized regressions, tree-based, and neural networks – as well as *more* complete than the *best* ML model overall (the horizontal lines). Over all subjects, the economic model is more complete for 59.7% and this fraction increases monotonically from 38.8% for subjects in the bottom quartile of consistency scores to 72.2% for subjects in the top quartile, who are (almost) perfectly rationalizable by the economic model. For those who are generally consistent with GARP and FOSD, there is little room for improving the prediction of the economic model. We note additionally that RDU is not less restrictive than most ML models, designed specifically for prediction, so its higher individual-level completeness indicates that it is better tuned to capture the heterogeneous demand behaviors of subjects.

[2]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)).

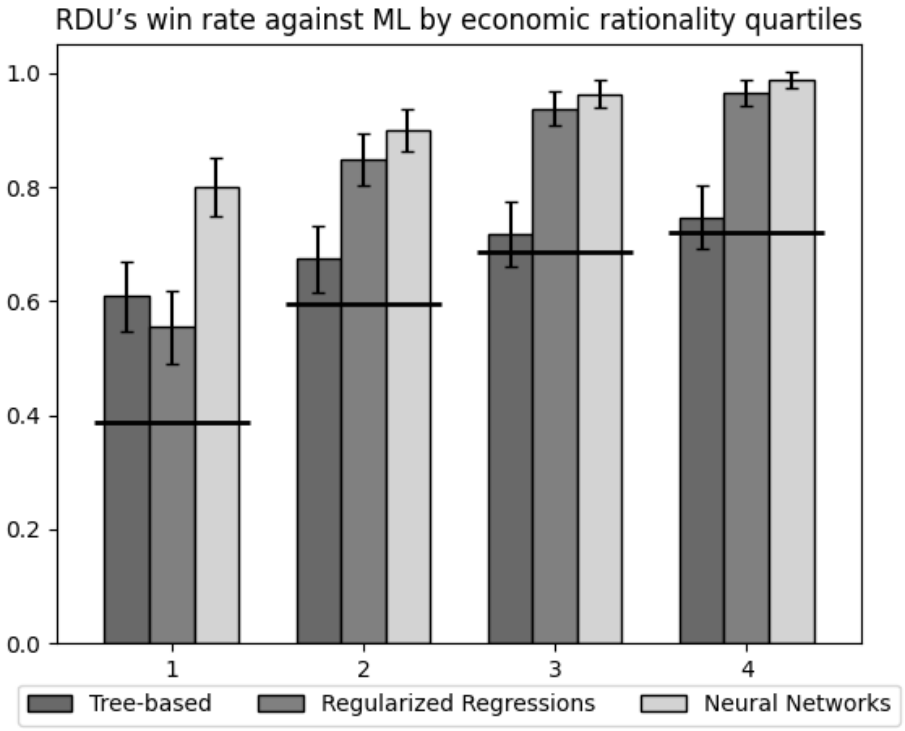


Figure 1: The fraction of subjects for whom RDU is more complete than the most complete regularized regressions, tree-based, and neural networks model, as well as more complete than the best ML model overall (the horizontal lines), quartiles of consistency scores with GARP and FOSD (Nishimura et al. (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. The closer it is to 1, the smaller the perturbation of budget lines required to remove all violations and thus the closer the data are to satisfying GARP and FOSD. The quartiles are [0, 0.831), [0.831, 0.950), [0.950, 0.988) and [0.988, 1).

Importantly, we also find that EUT outperforms RDU – EUT is more complete (by a small margin) for more subjects and more restrictive (because it is nested).[3] In Appendix Figure A.1, we present the results for EUT instead of RDU and the histograms appear nearly identical. Much of the experimental and behavioral literature on decisions under risk is directed towards finding violations of EUT. But EUT is part of the core of economics; it is not something that one can or should abandon lightly, even as a matter of parsimony. We interpret our results as a 'victory' for EUT which is foundational to so much of economics. This is of course based on one set of experimental results so we explore this and other themes in work-in-progress based on extensions of the present experimental design.

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

[3]This is consistent with the results of Choi et al. (2007a) that the parameter estimates vary dramatically across subjects, yet about half of the subjects are well-approximated by preferences consistent with EUT.

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 ML techniques to enhance economic models – theoretical and empirical. We will not attempt to review this large and growing literature. Mullainathan and Spiess (2017), Athey (2018) and Kleinberg et al. (2018) provide an excellent, though now somewhat dated, overview/assessment of the contributions of ML to economics. Instead, we focus attention on some recent papers that compare standard economic models of individual decision-making to ML models. Fudenberg et al. (2020) provide a more thorough review of the particularly relevant papers to our study that the reader may wish to consult.

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. Peysakhovich and Naecker (2017) also report that the economic models of choice under risk are substantially outperformed by regularized regressions on the aggregated data but perform equally well when individual-level parameters are included.

Fudenberg et al. (2021) and Fudenberg et al. (2020) respectively develop the measures of completeness and restrictiveness, which we adopt here to evaluate a model's prediction accuracy and flexibility. Fudenberg et al. (2021) 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). Fudenberg et al. (2021) 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. Using the same experimental data, Fudenberg et al. (2020) 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.[4]

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 extreme 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

[4]We do not compare the performances of non-EUT models. As Dembo et al. (2021) show, data from three-dimensional budget sets – involving three states with three associated securities – provide a much stronger test in terms of power than data from two-dimensional budget lines used in this paper, especially of the various generalizations of EUT that differ widely by how they weaken the independence axiom (while maintaining ordering and monotonicity with respect to FOSD). In the case of three states, the prominent non-EUT models make a specific and quite extreme set of restrictions on the structure of the utility function, thus yielding a set of empirically testable restrictions on observed behavior. We are pursuing this in a separate paper using further experimental data.

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, 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

In this section, we define the key concepts that we refer to throughout the paper. We first explain the dual measures of *completeness* and *restrictiveness* (Fudenberg and Liang 2020), with which we evaluate the prediction *accuracy* and *flexibility* of a model. We then describe the RDU model and ML models that we evaluate. To economize on space, we relegate the description of the experimental design and procedures to Appendix B. The technical discussion on testing for consistency with GARP and FOSD is relegated to Appendix C.

3.1 Measures

In our preferred interpretation of the experiment, there are two equally probable 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 $\mathcal{D} := (\mathcal{B}^i, x^i)$ 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.

Following the terminology and notation of Fudenberg et al. (2020), a *predictive mapping* $f : \boldsymbol{B} \to [0, 1]$ is a map from budget lines into token shares. Mappings are evaluated using the mean-squared error (MSE) *loss function* $\ell : [0, 1] \times [0, 1] \to \mathbb{R}$ where $\ell(f(\mathcal{B}^i), x^i)$ is the error assigned to a predicted token share $f(\mathcal{B}^i)$ when the chosen token share is x^i, so the normalized maximum error per observation is 1. The expected prediction error for a mapping f is the expected loss

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

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

[5] More precisely, the data generated by an individual's choices are $\left\{(\bar{x}_1^i, \bar{x}_2^i, x_1^i, x_2^i)\right\}_{i=1}^{50}$, where $(\bar{x}_1^i, \bar{x}_2^i)$ are the endpoints of the budget line and (x_1^i, x_2^i) 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.

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 $f^*_\Theta = \arg\min_{f\in\mathcal{F}_\Theta} \mathcal{E}_P(f)$.[6]

In recent work, Fudenberg and Liang (2019) and Fudenberg et al. (2021) 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 an 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 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.

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

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

We consider three parameter-free mappings as naive benchmarks: (i) equal allocations ($x_1 = x_2$) consistent with infinite risk aversion, (ii) allocating all the tokens to the cheaper security (x_1 if $p_1 < p_2$ and x_2 if $p_1 > p_2$) consistent with risk neutrality, and (iii) equal expenditures ($p_1x_1 = p_2x_2$) consistent with logarithmic von Neumann-Morgenstern utility function $\log x_1 + \log x_2$.[7] The maximum completeness value among this three naive benchmarks is taken to be the reported completeness of a model for an *individual* subject.

We emphasize the term individual to highlight that we analyze prediction accuracy at the

[6] We use 10-fold cross-validation as an estimate of prediction accuracy. This procedure is standard, and we refer the interested reader to Fudenberg et al. (2021) for full details. All our experiments and programs for analysis are available upon request.

[7] These three naive benchmarks also correspond to simple rules of thumb, or heuristics: (i) the point on budget line that lies on the 45-degree line, (ii) the highest the endpoints of the budget line, and (iii) the middle of the budget line. Note that unlike aggregate-level choices, the individual-level choices of some subjects can be best predicted by one of the naive benchmarks.

individual level. Clearly, even a high level of consistency in the individual-level decisions does not imply that aggregate data are consistent. In fact, the considerable heterogeneity in subjects' behaviors entails that although behaviors are *individually* consistent, they are *mutually* inconsistent. Thus, any aggregate-level estimation of an economic model in inevitably misspecified because there is no utility function that pooled choices maximize (Afriat's Theorem).

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 μ denote a distribution over mappings from $\mathcal{F}_{\mathcal{M}}$. For any two mappings f and f', define the distance between the two functions as

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

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

$$r_\Theta = \mathcal{E}_\mu \left[\frac{d(\mathcal{F}_\Theta, f)}{d(f_n, f)} \right].$$

Like completeness, we use the three naive benchmarks above to calculate restrictiveness, and we take the maximum of these restrictiveness values to be the reported restrictiveness of a model. We use bootstrapped data to generate a sample of simulated subjects to calculate restrictiveness.

3.2 The economic model

We consider a preference ordering in which the indifference curves have a 'kink' at the 45-degree line, which is consistent with Quiggin's RDU model (Quiggin (1982)).[8] The rank-dependent utility function takes the form:

$$U(x_L, x_H) = \beta_L u(x_L) + \beta_H u(x_H),$$

where $\beta_L, \beta_H > 0$ are decision weights that sum to unity, (x_L, x_H) is a *rank-ordered* allocation with payoffs $x_L \leq x_H$, and $u(\cdot)$ is the Bernoulli index. The economic parameter vector θ is thus the

[8] 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 rank-dependent utility is the leading contender. See Diecidue and Wakker (2001) for a comprehensive discussion.

decision weights and the Bernoull index. The RDU formula can be expressed in terms of a probability weighting function w, which is increasing and satisfies $w(0) = 0$ and $w(1) = 1$, as follows:

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

The RDU formulation encompasses a number of non-EUT models and embeds EUT as a parsimonious and tractable special case when $\beta_L = \beta_H$ (since each state has an equal likelihood of occurring).[9] If $\beta_L > \beta_H$, interpreted as "pessimism", the indifference curves have a 'kink' at safe allocations, where $x_1 = x_2$, that lie 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] For each subject, we estimated the RDU model using a constant relative risk aversion (CRRA) specification and a constant absolute risk aversion (CARA) specification.[11] 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.

3.3 Machine learning models

We consider eight 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. To economize on space, we briefly describe, but do not fully cover, each model.[13]

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

[9] Another interpretation of this preference ordering is that it displays loss or disappointment aversion (Gul, 1991). In this interpretation, the safe allocation $x_1 = x_2$ is taken to be the reference point.

[10] We note that while we do make comparisons between EUT and non-EUT models of choice under risk, in addition to our main comparison to ML models, the two-dimensional test has relatively low power. As pointed out by Dembo et al. (2021), an experiment involving three states and three associated securities has a number of important advantages in comparing EUT to non-EUT alternatives over experiments involving two states and two associated securities used here. Dembo et al. (2021) conclude that violations of EUT run deeper than violations of the Independence Axiom, thus challenging the most prominent non-EUT alternatives.

[11] For CARA, we assume $u(\cdot)$ takes the power form $u(x_s) = x_s^{1-\rho}/(1-\rho)$ where $\rho \geq 0$ is the Arrow-Pratt measure of relative risk aversion. For CRRA, 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 $\theta = (\beta, \rho)$ for CRRA and $\theta = (\beta, \gamma)$ for CARA.

[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[,]"

[13] We will not attempt to review the large and growing literature on machine learning. We provide references to seminal papers to refer to specific models. Hastie et al. (2009) and Daumé (2017) provide an in-depth treatment that the reader may wish to consult.

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, with two sets of covariates: (i) The set of budget lines $\boldsymbol{B}$ for each subject and subject indicator variables, a specification denoted simply as LASSO. (ii) A specification denoted LASSO+ also including interactions between $\boldsymbol{B}$ and the subject indicator variables. The former is nested within the latter. Second, we consider kernel ridge regression (KRR) (Hoerl and Kennard (1970)), which penalizes using the L_2 norm, using the latter specification with interactions between $\boldsymbol{B}$ and the subject indicator variables. Note that LASSO+ and KRR have the same number of parameters. The parameter vector θ for regularized regressions models is a linear coefficient vector.

Tree-based Unlike the linear relationship assumed in regularized regression, tree-based models divide the set of budget lines $\boldsymbol{B}$ 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 subset of budget lines, the model considers a further binary partition that minimizes the size-weighted error of both partitions. 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 two extensions, known more broadly as model trees (Quinlan et al. (1992)), change the estimated model from a sample mean to a linear regression (Linear), and support vector regression (SVR) with a normal radial basis function. The former is more familiar to economists, whereas the latter considers a nonlinear case that minimizes error whilst remaining as "flat" as possible. Mean is nested in Linear and SVR.

The last tree-based model, the random forest model (RF) averages the decision rules of multiple standard decision trees. Each tree is given only a subset of data, and is generally seen as an improvement over singular decision trees (Breiman (2001)). Trees are regularized by limiting the number of partitions and the minimum number of observations allowed within a partition. RF additionally regulates the number of trees. Because each tree is trained on a strict subset of data, 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 feed-forward neural networks, repeatedly transform input data through nonlinear functions of nonlinear functions, and subsequently aggregated together into a single output. Each "layer" of functions can be split into two stages: The input data is first transformed using a prespecified "activation function," such as $\max(x, 0)$ or $\tan(x)$. Then, these transformations are weighted and summed, to form a new vector of data known as "hidden units." These hidden units then become the input data to a new layer. To generate a demand prediction, the final layer output is one-dimensional. Within this space, specifying larger-dimensional ("wider") vectors of hidden units within each layer reduces restrictiveness. We implement a neural network using parameter values from Zhao et al. (2020), using cross-validation to determine the optimal number of layers and hidden units per layer. Conditional on the network "architecture," the parametrization θ of neural networks can be characterized by the weights.

4 Results

The experiment allows us to analyze behavior at the level of the individual subject and to testing whether choices are consistent with the primary axioms of revealed preference. 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, and the next column reports the win rate of RDU against each model (that is, the fraction of subjects for whom RDU is more complete). The next two blocks of four columns report the win rate of RDU 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 report the results for each regularized regression and tree-based model, respectively.

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

			RDU's win rate against ML by rationality quartiles				Absolute completeness difference between RDU and ML by rationality quartiles				
Panel A: RDU and ML model classes	Average completeness	RDU's win rate against ML	1st	2nd	3rd	4th	1st	2nd	3rd	4th	Restrictiveness
RDU	89.8% [89.1%, 90.5%]	-	-	-	-	-	-	-	-	-	23.4%
Regularized Regressions	82.6% [81.7%, 83.5%]	82.6%	55.4%	84.9%	93.8%	96.6%	0.9%	5.6%	7.5%	14.8%	27.4%
Tree-based Models	89.5% [88.8%, 90.2%]	68.6%	60.8%	67.4%	71.7%	74.7%	1.4%	0.7%	0.8%	1.0%	14.8%
Neural networks	79.7% [78.4%, 80.8%]	91.2%	79.6%	90.0%	95.8%	99.6%	5.7%	9.0%	9.2%	16.8%	20.4%
Panel B: Regularized regressions											
LASSO	78.0% [76.9%, 79.0%]	88.6%	71.3%	91.2%	93.8%	98.3%	5.0%	9.2%	11.8%	21.6%	29.3%
LASSO+	78.3% [77.2%, 79.3%]	87.3%	65.8%	91.2%	94.2%	98.3%	3.4%	9.1%	11.9%	21.7%	27.1%
Ridge	82.1% [81.2%, 82.9%]	88.0%	67.9%	88.3%	96.3%	99.6%	1.6%	6.1%	8.1%	15.2%	27.1%
Panel C: Tree-based models											
Mean	86.9% [86.1%, 87.7%]	84.4%	77.1%	88.3%	85.8%	86.5%	2.7%	3.7%	2.7%	2.5%	17.4%
Linear	83.6% [82.5%, 84.6%]	86.5%	80.8%	85.8%	87.1%	92.4%	10.1%	5.7%	4.4%	4.5%	7.6%
SVR	86.0% [85.2%, 86.8%]	88.5%	80.0%	90.4%	87.5%	96.2%	3.6%	3.9%	3.3%	4.4%	15.1%
RF	88.4% [87.7%, 89.1%]	79.8%	69.6%	78.7%	80.4%	90.7%	0.3%	1.6%	1.6%	2.1%	16.8%

Three main insights arise from Panel A of Table 1 about the prediction accuracy (completeness) and flexibility (restrictiveness) of the economic model 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 RDU is comparable to the completeness of the tree-based models (achieving 89.8% and 89.5% 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 82.6% and 79.5%, respectively). Furthermore, RDU's completeness win rate increases from 68.6% against tree-based models to 82.6% against regularized regression

models and to 91.2% against neural networks.

- Second, the win rate of RDU 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 RDU 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 RDU does not achieve a large improvement in completeness compared to tree-based models, it is substantially more restrictive (23.4% compared to only 14.8%). Moreover, the restrictiveness of RDU is comparable to the restrictiveness of the regularized regression models and neural networks (achieving restrictiveness of 27.4% and 20.4%, respectively), but these ML models are significantly less complete than RDU.

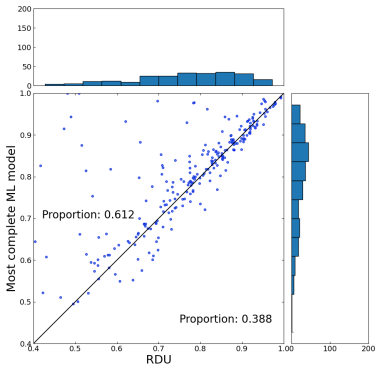
In Appendix Table A.1, we present near-identical results with EUT instead of RDU. Recall that the RDU model reduces to EUT when $\beta_L = \beta_H$ (since each state has an equal likelihood of occurring). EUT is therefore more restrictive than RDU (26.2% compared to 23.4%) and it is only moderately less restrictive than the regularized regression models (26.2% compared to 27.4%). Table 2 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. Panels B and C report the results assuming $u(\cdot)$ takes the CRRA and CARA specifications, respectively.[14] The main insights that arise from Table 2 are that the average completeness of EUT is (exactly) the same as the completeness of RDU but it is more restrictive. Also worthy of note is that EUT's overall win rate against RDU is above 50 percent but it is monotonically decreasing by consistency quartiles from the mid-high 60s in the bottom quartile to high 30s to low 40s in the top quartile. The relative improvement of EUT over RDU is essentially zero under both CARA and CRRA in all consistency quartiles. We consider this a 'victory' for the economic models, especially for the instrumental characterization of EUT.

[14] CARA is the more complete specification under EUT and RDU for 44.5% of our subjects whereas CRRA is the more complete under both for 28.6%. CARA is the more complete specification only under RDU for 15.8% of our subjects and it is the more complete specification only under EUT for 11.2%.

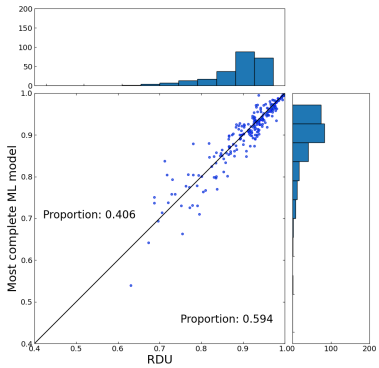
Table 2: The completeness and restrictiveness of EUT and RDU

	Average	EUT win rate	EUT's win rate against RDU by rationality quartiles				Absolute completeness difference between EUT and RDU by rationality quartiles				
Panel A: EUT and RDU	completeness	against RDU	1st	2nd	3rd	4th	1st	2nd	3rd	4th	Restrictiveness
EUT	89.8% [89.1%, 90.5%]	-	-	-	-	-	-	-	-	-	26.2%
RDU	89.8% [89.1%, 90.5%]	57.0%	67.5%	66.1%	51.3%	43.0%	0.5%	0.2%	-0.4%	-0.3%	23.4%
Panel B: CRRA Only											
EUT CRRA	89.4% [88.7%, 90.1%]	-	-	-	-	-	-	-	-	-	25.3%
RDU CRRA	89.4% [88.7%, 90.1%]	51.4%	65.4%	59.0%	43.3%	37.6%	0.6%	0.2%	-0.5%	-0.3%	23.1%
Panel C: CARA Only											
EUT CARA	89.1% [88.3%, 89.8%]	-	-	-	-	-	-	-	-	-	27.4%
RDU CARA	89.1% [88.3%, 89.8%]	56.1%	64.6%	64.9%	55.0%	39.7%	0.4%	0.1%	-0.3%	-0.2%	23.9%

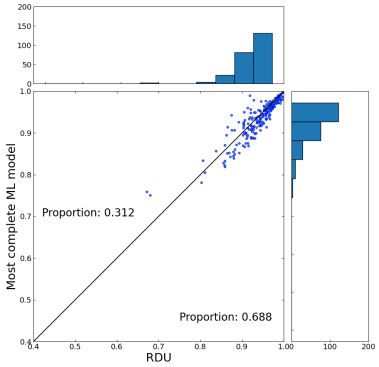
We can see this in greater detail in the four panels of Figure 2 below. We can see the comparison of completeness between RDU and the most complete ML model in greater detail in the four panels of Figure 2 below corresponding to the quartiles of the consistency of the individual-level data with GARP and FOSD. For each subject, the horizontal axis in each panel shows the completeness of RDU and the vertical horizontal axis shows the completeness of the best ML model. On each axis, we provide a histogram that shows the distribution of the completeness scores for each model. We first note that there are relatively few extreme differences in completeness, as seen by the absence of observations in the upper left and lower right corners of each panel, but there are few observations high above the diagonal in the bottom consistency quartile. We note additionally a monotonic shift towards the upper right corner by consistency quartiles, indicating greater completeness of both models as individual choices are more consistent. The fraction of observations below the diagonal (subjects for which RDU is the most complete model) increases by consistency quartile, and the distribution of completeness is higher for ML in the first panel (Kolmogorov-Smirnov test, $p = 0.018$) and higher for RDU in the fourth panel Kolmogorov-Smirnov test, $p = 0.004$). In Appendix Figure A.2, we again present near-identical results with EUT instead of RDU.



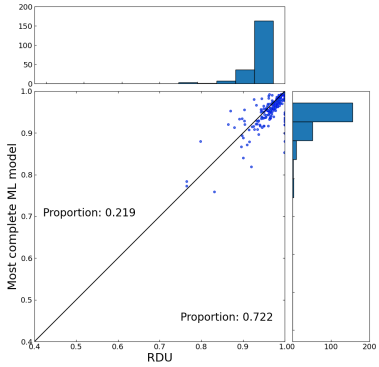
(a) Quartile 1



(b) Quartile 2



(c) Quartile 3



(d) Quartile 4

Figure 2: Scatterplot of completeness of RDU and the most complete machine learning model by rationality quartile.

5 Conclusion

We employ graphical representations of budget sets over bundles of state-contingent commodities, rather than discrete choices. This allows for the collection of a very rich individual-level data set. Our analysis begins by applying revealed preference tests to determine whether the observed choices are consistent with the axioms on which economic theory is based. We are able to provide a more precise comparison of the completeness of the economic model against a variety of ML models because we have measured the completeness of the different models at the individual level. Our main result is that the standard EUT model and the RDU model equally outperform all ML models, and by a wider margin the more consistent individual choices are with an underlying preference ordering. We consider this a victory for the economic models, especially for EUT as it is nested in RDU and thus more restrictive.

The experimental and analytical techniques serve as a foundation for comparing the completeness and restrictiveness of different models in more complex scenarios. We are already studying choices over three-dimensional budget sets, where different non-EUT models make a specific and quite

extreme set of restrictions on the structure of the utility function, thus yielding a set of empirically testable restrictions on observed behavior. Another promising direction is to study choice under ambiguity using the data of Ahn et al. (2014). The goal of this work is to generate analogous rigorous individual-level tests of the predictions of models of decision-making under ambiguity.[15] We are also studying within-subjects behavior across different treatments, for example, involving two-dimensional and three-dimensional budget sets. Clearly, the decision problem with two securities is a sub-problem of the decision problem with three securities and, if the subject has stable preferences, then economic models should predict choices across treatments.

References

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

Athey, S. (2018): "The impact of machine learning on economics," in *The economics of artificial intelligence: An agenda*, University of Chicago Press, 507–547.

Athey, S. and G. W. Imbens (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685–725.

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

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.

Daumé, H. (2017): "A course in machine learning: Hal Daumé III," .

Deaton, A. and J. Muellbauer (1980): "An almost ideal demand system," *The American economic review*, 70, 312–326.

Dembo, A., S. Kariv, M. Polisson, and J. K.-H. Quah (2021): "Ever Since Allais," Tech. rep., IFS Working Paper.

Diecidue, E. and P. P. Wakker (2001): "On the intuition of rank-dependent utility," *Journal of Risk and Uncertainty*, 23, 281–298.

Fudenberg, D., W. Y. Gao, and A. Liang (2020): "Quantifying the restrictiveness of theories," *Available at SSRN 3580408*.

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2021): "Measuring the Completeness of Economic Models," *Journal of Political Economy (forthcoming)*.

[15] To the best of our knowledge, only Peysakhovich and Naecker (2017) study choices under ambiguity, using standard discrete choices. They find that, unlike under risk, the economic models of decision-making under ambiguity do not predict individual choices as well as ML models.

Fudenberg, D. and A. Liang (2019): "Predicting and understanding initial play," *American Economic Review*, 109, 4112–41.

Gul, F. (1991): "A theory of disappointment aversion," *Econometrica: Journal of the Econometric Society*, 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.

Kahneman, D. and A. Tversky (1979): "Prospect theory: An analysis of decision under risk," *Econometrica*, 47, 363–391.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018): "Human decisions and machine predictions," *The quarterly journal of economics*, 133, 237–293.

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

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.

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 Reviewing (forthcoming)*.

Quiggin, J. (1982): "A theory of anticipated utility," *Journal of Economic Behavior & Organization*, 3, 323–343.

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: Journal of the Econometric Society*, 945–973.

——— (1983): "Non-parametric tests of consumer behaviour," *The review of economic studies*, 50, 99–110.

Zhao, C., S. Ke, Z. Wang, and S.-L. Hsieh (2020): "Behavioral neural networks," *Available at SSRN 3633548*.

Appendix A Additional figures and tables

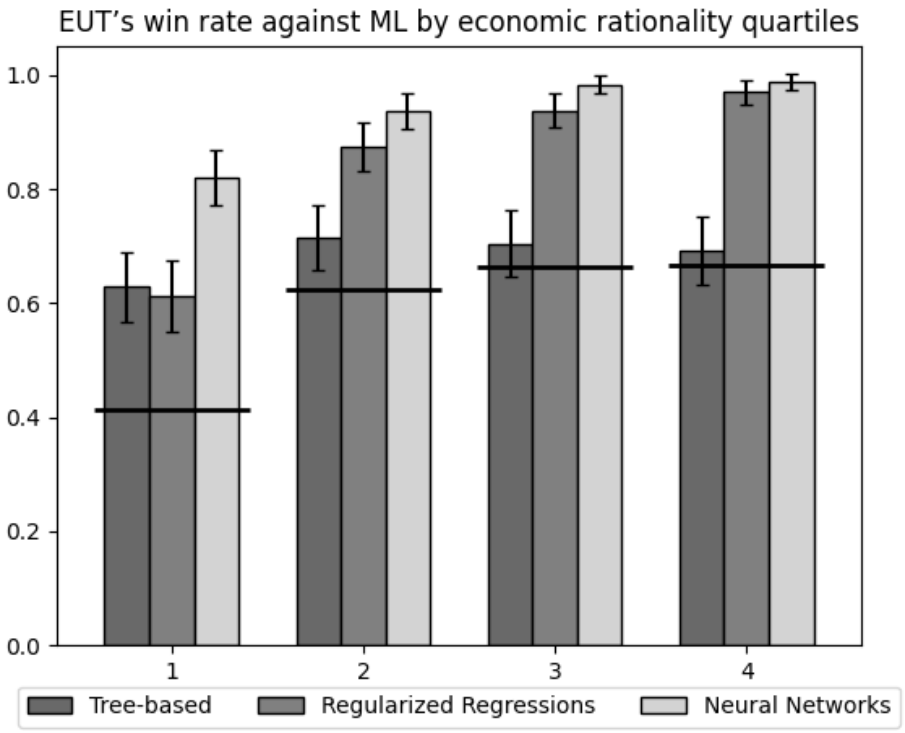
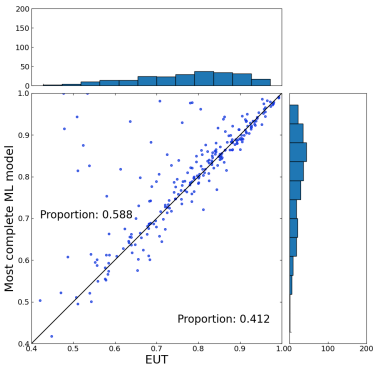


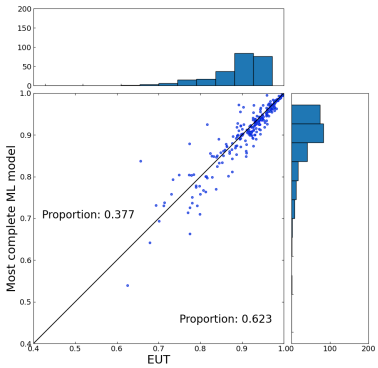
Figure A.1: The fraction of subjects for whom EUT is more complete than the most complete regularized regressions, tree-based, and neural networks model, as well as more complete than the best ML model overall (the horizontal lines), quartiles of consistency scores with GARP and FOSD (Nishimura et al. (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. The closer it is to 1, the smaller the perturbation of budget lines required to remove all violations and thus the closer the data are to satisfying GARP and FOSD. The quartiles are [0, 0.831), [0.831, 0.950), [0.950, 0.988) and [0.988, 1).

Table A.1: The completeness and restrictiveness of EUT and ML models

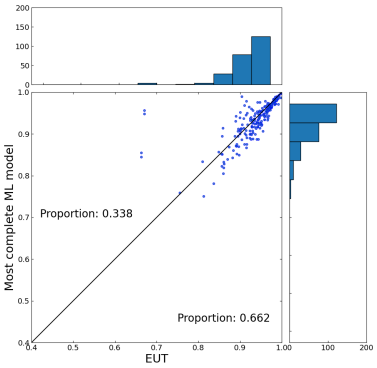
	Average	EUT's win rate	EUT's win rate against ML by rationality quartiles				Absolute completeness difference between EUT and ML by rationality quartiles				
Panel A: EUT and ML model classes	Completeness	against model	1st	2nd	3rd	4th	1st	2nd	3rd	4th	Restrictiveness
EUT	89.8% [89.1%, 90.5%]	-	-	-	-	-	-	-	-	-	26.2%
Regularized Regressions	82.6% [81.7%, 83.5%]	82.6%	61.3%	87.4%	93.8%	97.0%	1.4%	5.8%	7.2%	14.6%	27.4%
Tree-based Models	89.5% [88.8%, 90.2%]	68.6%	62.9%	71.5%	70.4%	69.2%	0.9%	0.9%	0.4%	0.7%	14.8%
Neural networks	79.7% [78.4%, 80.8%]	91.2%	82.1%	93.7%	98.3%	98.7%	6.1%	9.2%	8.8%	16.5%	20.4%
Panel B: Regularized regressions											
LASSO	78.0% [76.9%, 79.0%]	88.6%	77.5%	94.1%	96.7%	97.5%	5.4%	9.4%	11.5%	21.3%	29.3%
LASSO+	78.3% [77.2%, 79.3%]	87.3%	71.7%	94.1%	95.8%	97.5%	3.9%	9.3%	11.5%	21.5%	27.1%
Ridge	82.1% [81.2%, 82.9%]	88.0%	73.3%	91.6%	97.5%	99.2%	2.1%	6.3%	7.7%	14.9%	27.1%
Panel C: Tree-based models											
Mean	86.9% [86.1%, 87.7%]	84.4%	79.2%	89.1%	86.3%	80.2%	3.1%	3.9%	2.4%	2.2%	17.4%
Linear	83.6% [82.5%, 84.6%]	86.5%	83.3%	87.4%	87.1%	90.7%	10.6%	5.9%	4.1%	4.3%	7.6%
SVR	86.0% [85.2%, 86.8%]	88.5%	80.8%	88.7%	92.5%	94.5%	4.0%	4.1%	3.0%	4.1%	15.1%
RF	88.4% [87.7%, 89.1%]	79.8%	73.3%	81.2%	82.9%	86.1%	0.7%	1.7%	1.2%	1.9%	16.8%



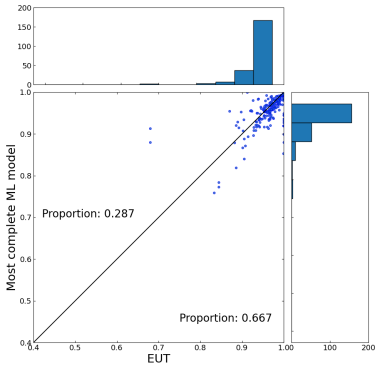
(a) Quartile 1



(b) Quartile 2



(c) Quartile 3



(d) Quartile 4

Figure A.2: Scatterplot of completeness of EUT and the most complete machine learning model by rationality quartile.

Appendix B The experiment

Choi et al. (2007b) developed an experimental graphical interface that allows subjects to make numerous choices over a wide range of budget sets, and this yields a rich dataset that is well-suited to analysis at the level of the individual subject. With the interface, subjects see on a computer screen a geometrical representation of a standard consumer decision problem – the selection of a bundle of commodities from a standard budget set – and choose allocations through a simple "point-and-click." The experiment consisted of 50 independent decision problems. In each decision problem, a subject was asked to allocate tokens (the experimental currency) 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.[16]

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 randomly selected one of the accounts, x or y, determined at random and equally likely. Each subject only 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 dataset is comprised of nearly a thousand subjects from several studies including the (symmetric) data collected by Choi et al. (2007a) and data from identical experiments with different subject pools collected by Zame et al. (2020) and Cappelen et al. (2021), as well as new data from identical experiments.[17] In all of these experiments, the individual-level data consist of 50 decision problems.[18] See Choi et al. (2007b) and Choi et al. (2007a) for an extended description of the experimental interface.[19]

[16]The budget lines selected for each subject in their 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 to move the pointer on the computer screen to the desired allocation. Choices were restricted to allocations on the budget constraint, so that subjects could not violate budget balancedness.

[17]Choi et al. (2007a) studied a symmetric treatment, in which the two accounts were equally likely and two asymmetric treatments in which one of the accounts was always selected with probability 1/3 and the other account was selected with probability 2/3.

[18]We do not include the data of Choi et al. (2014) which consist of 25, rather than 50, decision problems. The datasets of Choi et al. (2007a) and Choi et al. (2014) have been analyzed in many papers, including Halevy et al. (2018), Polisson et al. (2020), and De Clippel and Rozen (2021), among others.

[19]The experimental platform is applicable to many other types of individual choice problems. Ahn et al. (2014) extended the earlier experimental work of Choi et al. (2007a) in settings with risk (known probabilities) to settings with ambiguity (unknown probabilities). Fisman et al. (2007), Fisman et al. (2015a), Fisman et al. (2015b), Fisman et al. (2017) and Li et al. (2017), Li et al. (2022) employ a similar experimental methodology to study social preferences across a number of different samples, including a nationally representative sample.

Appendix C Revealed preference tests

The most basic question to ask about choice data is whether it is consistent with individual utility maximization. We thus want to relate the out-of-sample prediction accuracy of the economic model, as well as of the ML models, to the consistency of individual behaviors with utility maximization. If budget sets are linear (as in our experiments), 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. Because GARP provides an exact test of utility maximization – either the data satisfy GARP or they do not – but individual choices frequently involve at least some errors, we assess how nearly individual choice behavior complies with GARP by using Afriat (1972) critical cost efficiency index (CCEI), which measures the fraction by which each budget constraint must be shifted in order to remove all violations of GARP. By definition, the CCEI is between 0 and 1: indices closer to 1 mean the data are closer to perfect consistency with GARP and hence to perfect consistency with utility maximization.

But not any consistent preference ordering is admissible. Clearly, choices can be consistent with GARP yet fail to be reconciled with any utility function that is normatively appealing given the decision problem at hand. Given the two states are equally likely, allocating fewer tokens to the cheaper security ($x_s > x_{s'}$ when $p_s < p_{s'}$) is a violation of monotonicity with respect to FOSD. Violations of FOSD are errors – the failure to recognize that some allocations yield payoff distributions with unambiguously lower returns.[20] To test whether choice behavior satisfies GARP and FOSD (for a given subject), we combine the actual data from the experiment and the mirror-image data and compute the CCEI for this combined data set. By definition, the CCEI score for the combined data set consisting of 100 observations can be no bigger than the CCEI score for the actual data. Relying on Nishimura et al. (2017), Polisson et al. (2020) show that when states are equiprobable (as in our experiment), the CCEI score for the combined data set is a measure of consistency with GARP and FOSD for each subject because a well-behaved utility function is monotone with respect to FOSD if and only if it is symmetric.[21]

[20] Almost all decision-theoretic models that have been proposed as alternatives to EUT of which we are aware obey monotonicity with respect to FOSD, including RDU (Quiggin 1982, 1993), Weighted Expected Utility (Dekel 1986; Chew 1989), and CPT (Tversky and Kahneman (1992)). As noted by Quiggin (1990), Wakker and Tversky (1993) and Starmer (2000) prominent non-EUT models, including Prospect Theory (Kahneman and Tversky (1979)), were amended to avoid violations of FOSD.

[21] Clearly, any decision to allocate fewer tokens to the cheaper security (positions along the shorter side of the budget line relative to the 45-degree line) will necessarily generate a simple violation of the Weak Axiom of Revealed Preference (WARP) involving its mirror-image decision.

Additional References

Afriat, S. N. (1972): "Efficiency estimation of production functions," *International economic review*, 568–598.

Cappelen, A. W., S. Kariv, E. Ø. Sørensen, and B. Tungodden (2021): "The Development Gap in Economic Rationality of Future Elites," *Unpublished paper*.

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

De Clippel, G. and K. Rozen (2021): "Bounded rationality and limited data sets," *Theoretical Economics*, 16, 359–380.

Fisman, R., P. Jakiela, and S. Kariv (2015a): "How did distributional preferences change during the great recession?" *Journal of Public Economics*, 128, 84–95.

——— (2017): "Distributional preferences and political behavior," *Journal of Public Economics*, 155, 1–10.

Fisman, R., P. Jakiela, S. Kariv, and D. Markovits (2015b): "The distributional preferences of an elite," *Science*, 349.

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

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

Li, J., L. P. Casalino, R. Fisman, S. Kariv, and D. Markovits (2022): "Experimental evidence of physician social preferences," *Proceedings of the National Academy of Sciences*, 119, e2112726119.

Li, J., W. H. Dow, and S. Kariv (2017): "Social preferences of future physicians," *Proceedings of the National Academy of Sciences*, 114, E10291–E10300.

Quiggin, J. (1990): "Stochastic dominance in regret theory," *The Review of Economic Studies*, 57, 503–511.

Tversky, A. and D. Kahneman (1992): "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and uncertainty*, 5, 297–323.

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 (2020): "Linking Social and Personal Preferences: Theory and Experiment," *Unpublished paper*.