What can the demand analyst learn from machine learning?[*]

Keaton Ellis, Shachar Kariv and Erkut Ozbay[†]

September 24, 2023

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 et al. (2022a,b), 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.

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

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), derives the associated demand function, and fits it to data using some econometric technique. There is a wide variety of formats to this economic approach, ranging from nonparametric to semiparametric to parametric methods. The estimated preference parameters can then be used to extrapolate and 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* by Fudenberg et al. (2022b,a). 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 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 956 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]

[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.

[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)).

For each subject, we calculate the completeness and restrictiveness of Expected Utility Theory (EUT). 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). EUT, as well as almost all models that generalize EUT, embody ordering (completeness and transitivity) and monotonicity.

Figure 1 depicts our main result. The horizontal axis presents quartiles of consistency scores with GARP and FOSD. Note that the quartiles are over the distribution of subjects scores. The consistency score intervals for the quartiles are [0, 0.831), [0.831, 0.950), [0.950, 0.988) and [0.988, 1). The vertical axis indicates the fraction of subjects for whom EUT 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 65.4% and this fraction increases monotonically from 54.2% for subjects in the bottom quartile of consistency scores to 73.8% for subjects in the top quartile, who are (almost) perfectly rationalizable 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 EUT 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.

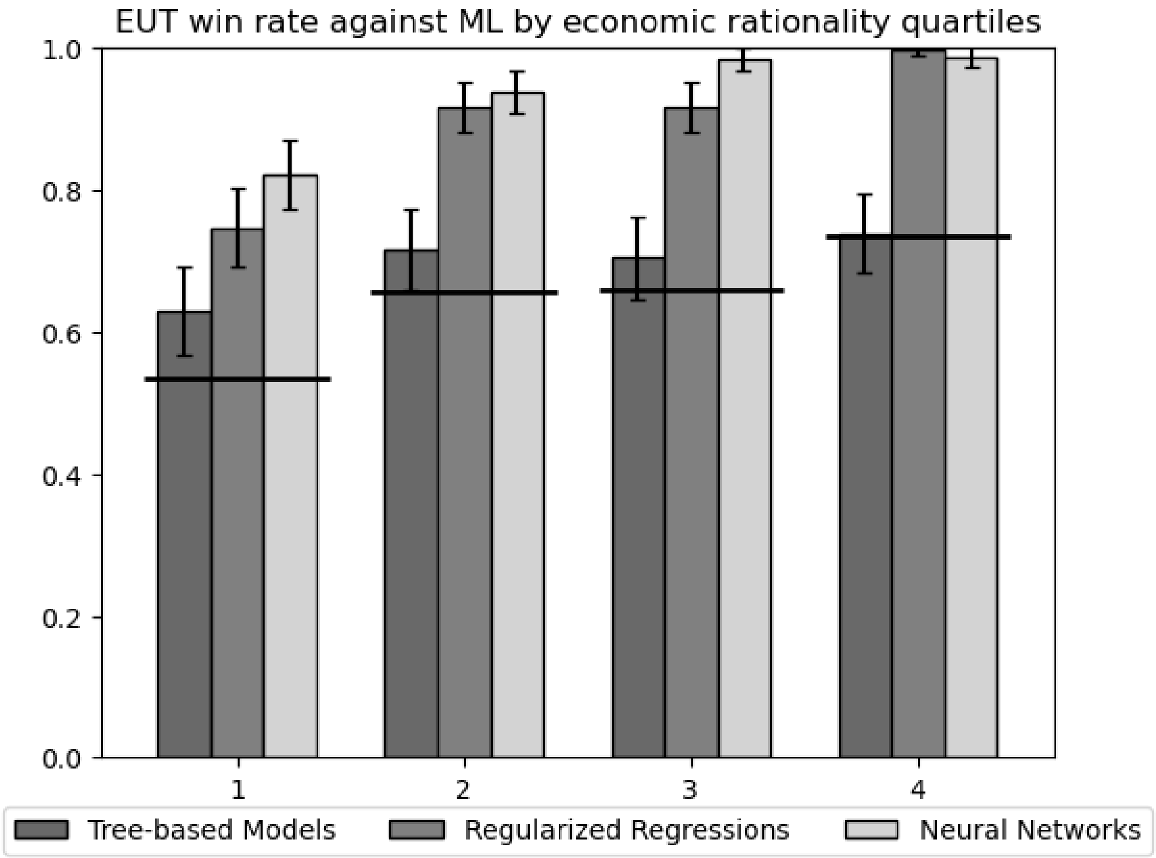


Figure 1: The fraction of subjects for whom EUT is more complete than the most complete regularized regressions, tree-based, and neural networks model, as well as more complete than the best ML model overall (the horizontal lines), quartiles of consistency scores with GARP and FOSD (Nishimura et al. (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).

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 RDU instead of EUT and the histograms appear nearly identical. Much of the experimental and behavioral literature

[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.

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 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. (2022a) 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. (2022b) and Fudenberg et al. (2022a) respectively develop the measures of completeness and restrictiveness, which we adopt here to evaluate a model's prediction accuracy and flexibility. Fudenberg et al. (2022b) 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. (2022b) 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. (2022a) 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 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.

[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.

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* by Fudenberg et al. (2022b,a), with which we evaluate the prediction *accuracy* and *flexibility* of a model. We then describe the EUT model, 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 equiprobable states of nature $s = 1, 2$ and an Arrow security for each state. Let $x_s \geq 0$ denote the demand for the security that pays off in state s and $p_s > 0$ denote the corresponding price. The budget set is then given by $\mathcal{B} = \{\mathbf{x} : \mathbf{p} \cdot \mathbf{x} = m\}$, where $\mathbf{x} = (x_1, x_2)$ is a demand allocation, $\mathbf{p} = (p_1, p_2)$ is a price vector and m is the endowment. We also define the *token share* of the security that pays off in one state to be the number of tokens payable in that state as a fraction of the sum of tokens payable in both states, and denote $x = x_1/(x_1 + x_2)$ to be the token share for the first state. Let $\mathcal{D} := (\mathcal{B}^i, x^i)$ 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. (2022a), 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

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

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

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

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

Since subjects see budget sets at most once, it is possible to construct a function from budget sets to demand that will achieve zero error, and thus we assume that $f^*(\mathcal{B}^i) = 0$. 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 error of a naive model is $\ell(x_{naive}, x) = (x_{naive} - x)^2$. If $x_{naive} \sim U[0, 1]$, then the expected squared error conditional on the actual choice x is:

$$\int_0^1 (x - \eta)^2 d\eta = \frac{1}{3}(1 - 3x + 3x^2)$$

We follow Fudenberg et al. (2022b) and use 10-fold cross-validation as an estimate of model expected error. For each subject, the data of 50 budget lines and associated choices is randomly partitioned into 10 "folds" of five observations. For each fold k, the other nine folds of 45 observations are used to estimate a model and obtain predictions for the budget lines in k. We then take the average MSE across the folds for an estimate of expected model error. The estimate of completeness is calculated by plugging in the cross-validation estimates of model, naive, and irreducible error. This estimate is shown to be consistent in Fudenberg et al. (2022b) and shown to be normally distributed around the true completeness value in Fudenberg et al. (2022a).

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

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

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

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

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 μ, 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 $x = \frac{x_1}{x_1+x_2}$ 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 drawn from μ. Each model is evaluated at the agent level, and the resulting within-sample errors are used to calculate restrictiveness.

3.2 The economic models

We consider preferences orderings consistent with classical expected utility (EUT). In this environment, the EUT utility function takes the form

$$U(x_1, x_2) = 0.5u(x_1) + 0.5u(x_2),$$

where $u(\cdot)$ is the Bernoulli index. For each subject, we estimated the EUT model using a constant relative risk aversion (CRRA) specification and a constant absolute risk aversion (CARA) specification.[6] 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.

We additionally consider a non-EUT behavioral model, which is consistent with Quiggin's rank-dependent utility (RDU) model (Quiggin (1982)).[7] The RDU function takes the form:

$$U(x_L, x_H) = (1 - w)u(x_L) + wu(x_H),$$

where $w \in (0, 1)$, (x_L, x_H) is a *rank-ordered* allocation with payoffs $x_L \leq x_H$, and $u(\cdot)$ is the Bernoulli index. The RDU formulation encompasses a number of non-EUT models and embeds EUT as a parsimonious and tractable special case when $w = 0.5$ (since each state has an equal likelihood of occurring).[8] If $w < 0.5$, 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).[9]

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

[6] For CRRA, 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 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.

[7] 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.

[8] 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.

[9] 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.

ML, and increasingly economics, literatures. We include multiple approaches because there is no declared 'winning' method.[10] 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 in the main text, we only briefly describe each model, elaborating in more detail in Appendix D.[11]

Regularized regressions Regularized regression, in its simplest form, assumes a linear relationship between outcomes and covariates, whose coefficient is estimated using 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. 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. 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 affects the degree to which the magnitudes of coefficients affect the objective function.

Tree-based Unlike the linear relationship assumed in regularized regression, 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. [12]

[10] 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[,]"

[11] 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.

[12] 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 Appendix D.

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 (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, (RF) averages the decision rules of multiple standard decision trees. Each tree is given a bootstrapped data set, and is generally seen as an improvement over singular decision trees (Breiman (2001)). Because each tree not trained on the original data set, there is no nesting and thus no restrictiveness or completeness guarantees between RF and the other tree-based models. Additionally, since trees are inherently nonparametric, they cannot be easily described by a parameter vector θ.

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

For a multilayer perceptron, the parameter values $W^{(i)}$ and $b^{(i)}$ are estimated, while the analyst has the freedom to choose the number of layers, the dimensions of each layer, the $\sigma^{(i)}$ functions, and a number of parameters associated with the estimation of $W^{(i)}$ and $b^{(i)}$. We use the layer count, layer dimension, and $\sigma^{(i)}$ values from Zhao et al. (2020).[13]

[13] See Appendix D for more detail.

4 Results

The experiment allows us to analyze behavior at the level of the individual subject and to test 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, as well the 95% confidence interval for average completeness, and the next column reports the win rate of EUT against each model (that is, the fraction of subjects for whom EUT is more complete).[14] The next two blocks of four columns report the win rate of EUT against each model and its absolute completeness difference by quartiles of the consistency score with GARP and FOSD. The right column reports the restrictiveness of each model. Panel A of Table A.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.

[14] We calculate the mean completeness confidence intervals by bias-corrected and accelerated bootstrapping, with 10000 resamples of size 956.

Table 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	89.3% [88.4%, 90.0%]	-	-	-	-	-	-	-	-	-	18.5%
Regularized Regressions	79.5% [77.8%, 80.6%]	88.5%	72.1%	91.6%	91.3%	99.2%	3.7%	7.8%	9.7%	17.8%	20.6%
Tree-based Models	89.1% [88.4%, 89.9%]	69.7%	62.9%	71.5%	70.4%	73.8%	-1.4%	0.9%	0.5%	0.6%	9.4%
Neural Networks	71.6% [68.7%, 73.7%]	94.2%	84.6%	95.4%	97.9%	99.2%	9.3%	14.6%	16.6%	30.5%	14.3%
Panel B: Regularized regressions											
Lasso	75.9% [74.2%, 76.9%]	92.2%	80.0%	95.0%	94.2%	99.6%	7.0%	11.8%	13.8%	21.1%	20.6%
OLS	70.2% [57.5%, 74.7%]	90.3%	75.8%	92.5%	93.8%	99.2%	11.2%	10.6%	15.8%	38.8%	20.6%
Ridge	70.6% [58.2%, 75.0%]	90.3%	75.8%	92.5%	93.8%	99.2%	11.1%	10.4%	15.4%	37.9%	20.6%
Panel C: Tree-based models											
Mean	86.6% [85.7%, 87.4%]	84.8%	79.2%	89.1%	86.3%	84.8%	3.0%	3.7%	2.2%	1.9%	12.3%
Linear	82.9% [81.6%, 84.0%]	87.1%	83.3%	87.4%	87.1%	90.7%	12.4%	6.1%	3.5%	3.4%	5.4%
SVR	85.7% [84.8%, 86.6%]	89.1%	80.8%	88.7%	92.5%	94.5%	4.1%	4.1%	2.7%	3.3%	10.7%
RF	88.0% [87.2%, 88.8%]	80.9%	73.3%	81.2%	82.9%	86.1%	0.5%	1.7%	1.2%	1.5%	11.9%

Three main insights arise from Panel A of Table 1 about the prediction accuracy (completeness) and model 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 EUT is comparable to the completeness of the tree-based models (achieving 89.3% and 89.1% of the feasible reduction in prediction error, respectively), but it is significantly more complete than regularized regression models and neural networks (achieving completeness of only 79.5% and 71.6%, respectively). Furthermore, EUT's completeness win rate increases from 69.7% against tree-based models to 88.5% against regularized regression models and to 94.2% against neural networks.

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

- Third, while EUT does not achieve a large improvement in completeness compared to tree-based models, it is substantially more restrictive (21.6% compared to only 11.0%). Moreover, the restrictiveness of EUT is comparable to the restrictiveness of the regularized regression models and neural networks (achieving restrictiveness of 23.8% and 16.9%, respectively), but these ML models are significantly less complete than EUT.

In Appendix Table A.1, we present near-identical results with RDU instead of EUT. Recall that the RDU model reduces to EUT when $w = 0.5$ (since each state has an equal likelihood of occurring). RDU is therefore less restrictive than EUT (21.6% compared to 19.3%) and it is also only moderately less restrictive than the regularized regression models (19.3% compared to 23.8%). Table 2 provides a population-level summary of our results comparing the economic models, EUT and RDU, in the same format as Table A.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.[15] The main insights that arise from Table 2 are that the average completeness of EUT is the same as the completeness of RDU but it is more restrictive. 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 low 40s in the top quartile. Despite this, the absolute 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.

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	89.3%	-	-	-	-	-	-	-	-	-	21.6%
	[88.4%, 90.0%]										
RDU	89.2%	60.6%	68.3%	66.5%	53.8%	53.6%	0.6%	0.2%	-0.2%	-0.2%	19.3%
	[88.3%, 89.9%]										
Panel B: CRRA Only											
EUT CRRA	88.8%	-	-	-	-	-	-	-	-	-	20.9%
	[88.0%, 89.6%]										
RDU CRRA	88.8%	53.0%	65.4%	59.0%	43.3%	44.3%	0.7%	0.2%	-0.4%	-0.3%	19.0%
	[87.9%, 89.6%]										
Panel C: CARA Only											
EUT CARA	88.6%	-	-	-	-	-	-	-	-	-	22.5%
	[87.8%, 89.4%]										
RDU CARA	88.5%	58.3%	65.4%	64.9%	56.3%	46.4%	0.5%	0.2%	-0.2%	-0.2%	19.7%
	[87.7%, 89.3%]										

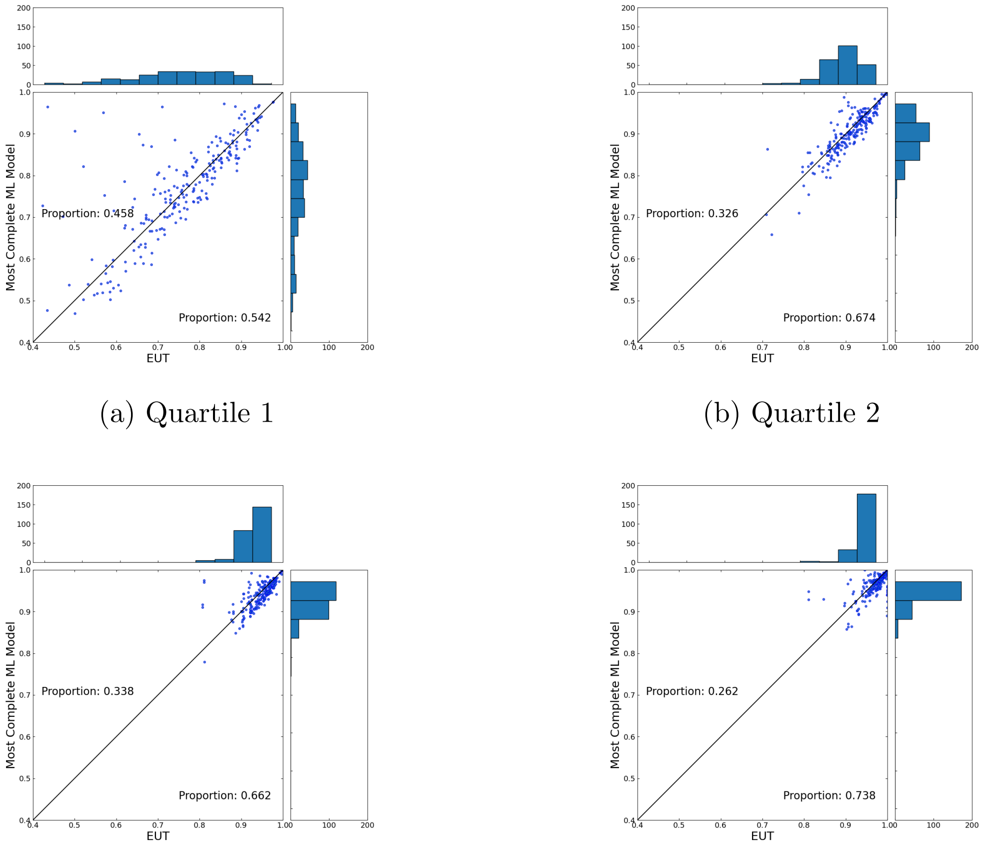
We can see this in greater detail in the four panels of Figure 2 below. We can see the comparison of completeness between EUT and the most complete ML model in greater detail in the four panels of Figure 2 below corresponding to the quartiles of the consistency of the individual-level data with GARP and FOSD. For each subject, the horizontal axis in each panel shows the completeness of EUT and the vertical horizontal axis shows the completeness of the best ML model. 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

[15] CARA is the more complete specification under EUT and RDU for 30.7% of our subjects whereas CRRA is the more complete under both for 42.0%. CARA is the more complete specification only under RDU for 12.3% of our subjects and it is the more complete specification only under EUT for 15.0%.

each panel, but there are few observations high above the diagonal in the bottom consistency quartile.[16] 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 EUT is the most complete model) weakly increases by consistency quartile, and the distribution of completeness is higher for EUT in all panels. Finally, we note a complementarity between machine learning models - of the 331 subjects for whom the most complete machine learning model is more complete than EUT, 249 (26.0%) have RDU as the second most complete model, above the other two classes of machine learning models.

In Appendix Figure A.2, we again present near-identical results with RDU instead of EUT.

[16] In Appendix G, we further examine these subjects and identify them primarily as having systematic deviations from FOSD (and for some, GARP). In Appendix H, we address statistical uncertainty regarding win rates. Of the 956 subjects in the data set, only 13 have significant differences in completeness.



(a) Quartile 1

(b) Quartile 2

(c) Quartile 3

(d) Quartile 4

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

As robustness checks, Appendix E replicates the analysis with the data from a similar budget line experiment with asymmetric probabilities. Appendix F replicates the analysis with a sample of hypothetical subjects who implement a log-utility with error.

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.

We view our analysis as a “best-case scenario” baseline, in an ideal environment to analyze the individual-level effectiveness of economic models relative to machine learning models. Hence, 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.[17] 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. Finally, we are interested in investigating whether auxiliary coviarates, such as sociodemographic information, can be of additional use over choice data for demand analysis.

[17] 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.

References

Afriat, S. N. (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.

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.

Bottou, L., F. E. Curtis, and J. Nocedal (2018): "Optimization methods for large-scale machine learning," *SIAM review*, 60, 223–311.

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

Bronars, S. G. (1987): "The power of nonparametric tests of preference maximization," *Econometrica: Journal of the Econometric Society*, 693–698.

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., 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.

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

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

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.

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.

Fudenberg, D., W. Y. Gao, and A. Liang (2022a): "How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories," *Available at SSRN 3580408*.

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2022b): "Measuring the completeness of economic models," *Journal of Political Economy*, 130, 956–990.

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

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.

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.

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

Quinlan, J. R. et al. (1992): "Learning with continuous classes," in *5th Australian joint conference on artificial intelligence*, World Scientific, vol. 92, 343–348.

Smola, A. J. and B. Schölkopf (2004): "A tutorial on support vector regression," *Statistics and computing*, 14, 199–222.

Starmer, C. (2000): "Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk," *Journal of economic literature*, 38, 332–382.

Sun, S., Z. Cao, H. Zhu, and J. Zhao (2019): "A survey of optimization methods from a machine learning perspective," *IEEE transactions on cybernetics*, 50, 3668–3681.

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

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

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.

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.*

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