# The Predictivity of Theories of Choice Under Uncertainty

Keaton Ellis, Shachar Kariv, and Erkut Ozbay[*]

Latest draft here

November 26, 2023

**Abstract**

Economic models are founded on parsimony and interpretability, which is achieved through axioms on choice behavior. We empirically evaluate the predictive accuracy of economic models of choice under risk and ambiguity, and the strength of their axiomatic foundations, using complementary methods of completeness (Fudenberg et al., 2022) and restrictiveness (Fudenberg et al., 2023), respectively. To better understand the tradeoff between the two concepts, we additionally relate their performance to machine learning models. We use budgetary choice environments with three dimensions to provide a strong test of axioms. We show that adding a third dimension of choice marginally reduces completeness of economic models, but significantly increases restrictiveness. Economic models are also more complete than machine learning models, and are significantly more restrictive. These results are robust to considering an environment of choice under ambiguity than choice under risk. We contrast these results with aggregate-level analysis, where we fit a single mapping of a model to the entire data set without identifying information about which subjects made which choice. All models have lower completeness, but machine learning models are more complete than economic models. Finally, we conduct a "group-level" analysis by providing the machine learning model with identifying information. All machine learning models outperform economic models, with higher gains for subjects least consistent with GARP. Overall, economic models capture the behavior of individual subjects well. However, machine learning models are able to effectively transfer information across subjects to improve prediction, at the cost of interpretability. Future work is needed to generate a better understanding of transferability of choice data.

[*] 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

In seminal recent and ongoing work, Fudenberg et al. (2022, 2023) propose a method to use machine learning techniques to evaluate and improve economic models. The key dual concepts in this regard are *completeness* and *restrictiveness* of a model. The completeness of a model is the fraction of the predictable variation in the data that the model captures. The restrictiveness of a model discerns completeness due to the "right" regularities by evaluating its completeness on arbitrary data. A more complete model better captures the regularities in the data, but the model might have enough flexibility to accommodate any regularity. An unrestrictive model is complete on any possible data, so the fact that it is complete on the actual data is uninstructive.

Ellis et al. (2022) evaluate the completeness and restrictiveness of the prominent economic models of choice under risk — as well as various families of machine learning models — using data from an economically important 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. (2007a,b) that allows for the collection of a rich individual-level data set. The large amount of data generated by this two-dimensional budget lines design allows us to apply the models to individual data rather than pooling data or assuming homogeneity across subjects.

More precisely, in Ellis et al. (2022), there are two equiprobable states of nature denoted by $s = 1, 2$ and two associated Arrow securities, each of which promises a dollar payoff in one state and nothing in the other. Let $\mathbf{x} = (x_1, x_2) \geq \mathbf{0}$ denote an allocation of securities, where $x_s$ denotes the number of units of security $s$. Without essential loss of generality, assume the individual's endowment is normalized to 1. An allocation $\mathbf{x}$ must satisfy the budget constraint $\mathbf{p} \cdot \mathbf{x} = 1$, where $\mathbf{p} = (p_1, p_2) \geq \mathbf{0}$ is the vector of security prices and $p_s$ denotes the price of security $s$.

In this paper, we extend the analysis of Ellis et al. (2022) by analyzing richer data using three-dimensional budget sets. In the case of three states, the axioms underpinning the economic models of choice under uncertainty make a specific and quite extreme set of restrictions on the structure of the utility function, thereby yielding a set of empirically testable restrictions on observed behavior. In particular, some axioms prescribe identical choice behavior in two dimensions, but differentiate in higher dimensional space. Separability, for instance, is trivially satisfied in two dimensions but not so in three. Like Ellis et al. (2022), for each subject, we have 50 observations $\left\{\mathbf{p}^i, \mathbf{x}^i\right\}_{i=1}^{50}$, where $\mathbf{p}^i$ denotes the $i$-th observation of the price vector $\mathbf{p} = (p_1, p_2, p_3) \geq \mathbf{0}$ and $\mathbf{x}^i$ denotes the associated allocation $\mathbf{x} = (x_1, x_2, x_3) \geq \mathbf{0}$.

We present two experiments with differing probability information using this richer data involving three states with three associated Arrow securities. First, we present an extensive elaboration to study choice under risk with *known* probabilities, a direct extension of Ellis et al. (2022). Second, we present an intensive elaboration that employs an analogous individual-level data set which similarly allows for a rigorous test of choice under ambiguity with *unknown* probabilities. We study choice under risk using the data of Dembo et al. (2021) and choice under ambiguity using the data of Ahn et al. (2014).

There is a variety of theoretical models of attitudes toward risk and ambiguity, but in this environment they all give rise to one of two specifications. The main specification we consider is a (generalized) "kinked" specification (Ahn et al., 2014) that builds on Rank-Dependent Utility (RDU) of Quiggin (1982). This specification is characterized by three parameters measuring attitudes towards risk, loss, and ambiguity and it can be derived as a special case of a variety of utility models in the literature, including Maxmin Expected Utility (MEU), Choquet Expected Utility (CEU), $\alpha$-Maxmin Expected Utility ($\alpha$-MEU), and Contraction Expected Utility.[1]

The other specification is "smooth" and it can be derived from the class of Recursive Expected Utility (REU) models. This specification is characterized by two parameters measuring attitudes towards risk and ambiguity.[2] Savage (1954) Subjective Expected Utility (SEU) model is a special cases of both specifications. Von Neumann and Morgenstern (1947) Expected Utility Theory (EUT) is a special case of the "kinked" specification.

Our results can be summarized in five headings.

**Adding dimension of choice** We show that there is only a slight decrease in completeness for EUT and RDU when adding a third dimension of choice. This is of particular interest because there is a large increase in restrictiveness for both models. The result is not driven by changes in consistency with the generalized axiom of revealed preference (GARP) . Hence, as discussed above, the assumptions underpinning economic models more strongly restrict demand in the higher dimensional space. Within the three dimensional environment, we show that EUT remains as complete as RDU, and is not significantly more restrictive. The results are consistent with a data generating process of EUT plus noise, but inconsistent with a data generating process of RDU. Overall, our results are consistent with the notion that violations of EUT do not manifest as violations of the independence axiom, which primarily distinguishes EUT and SEU from non-EUT models, but instead come in the form of violations of more core axioms such as GARP.

**Adding ambiguity** We observe two main results when comparing similar experiments of Risk and Ambiguity. First, we show that the completeness of SEU in the Ambiguity experiment is comparable to that of EUT in the Risk experiment. Thus, introducing ambiguity does not affect the completeness of the standard model. Second, we study a variety of models of choice under ambiguity, all of which nest SEU as a special case. We show that SEU is as complete as these

[1] See, MEU: Gilboa and Schmeidler (1989); CEU: Schmeidler (1989); $\alpha$-MEU: Ghirardato et al. (2004) and Olszewski (2007); Contraction Expected Utility: Gajdos et al. (2008).

[2] See, Ergin and Gul (2009), Klibanoff et al. (2005), Nau (2006), Giraud (2014), and Ahn (2008).

models and is marginally more restrictive. Our results again are consistent with the notion that violations of SEU run deeper than violations of the independence axiom.

**Comparing economic and machine learning models** We compare the completeness and restrictiveness of economic models to machine learning models. We use seven models across three families: regularized regressions, tree-based models, and neural networks. We show that the standard economic models of EUT and SEU are more complete than machine learning models for over 70% of all subjects. Economic models are also significantly more restrictive than machine learning models in both the Risk and Ambiguity experiments. Hence, the assumptions of economic models are "right", they have empirical content, and there are no regular choice patterns that economic models fail to implement.

**Comparing individual-level and aggregate-level analysis** We conduct a similar exercise assuming that the analyst does not have identifying information of which subject made each choice, which we interpret as misspecification cost. Thus, instead of fitting a separate mapping for each subject, we fit a single mapping of a model to the entire data set. We show that there is asymmetric misspecification cost between economic and machine learning models, and it is largest for economic models. Additionally, in contrast to individual-level analysis, aggregate-level machine learning models are more complete than aggregate-level economic models. These results emphasize that individual-level analysis should be conducted when possible. In particular, when analyzing economic model performance, doing so in an aggregate environment, where economic models are inherently misspecified, leads to incorrect conclusions about model quality. When identifying information is not available, however, economic models are so misspecified that one should use a machine learning model instead.

**Augmenting machine learning models with group-level data** Finally, we provide machine learning models with the entire data set, including identifying information. This is a trivial exercise for machine learning models, but not so for economic models. We show that these group-level machine learning models are more complete than any other model considered. Compared to individual-level models, the gains are largest for subjects least consistent with GARP. These results show that there is relevant between-subject information, and that group-level machine learning models can effectively incorporate this information.

Overall, our results show a tradeoff between performance and interpretability, and an analyst's model of choice should be dependent on analysis objectives, data availability, and potentially rules on data usage. Machine learning models provide maximal performance, but individual-level economic models provide second-best performance while additionally providing for easy interpretation. Additionally, our results indicate that there is little information that the standard models of

EUT and SEU do not incorporate within subject, but there is room for further research regarding between-subject analysis.

In Section 2, we discuss related literature. Section 3 discusses the experimental environment, the measures used to evaluate model quality, and economic and machine learning models. Section 4 discusses the power of the experimental design. Section 5 discusses the individual-level and aggregate-level results, and Section 6 discusses the group-level machine learning model results. Section 7 concludes.

## 2 Related Literature

Our paper contributes to the body of work that seeks to use machine learning techniques to evaluate and enhance economic models – theoretical and empirical. Peysakhovich and Naecker (2017) compare the performance 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" when 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 and Liang (2019) conduct a similar exercise, focusing on evaluating model prediction power of initial play in $3 \times 3$ matrix games. They find that the best model in their space is a hybrid model, which splits its prediction between level-1 thinking with curvature in utility and a Poisson cognitive hierarchy model. Agrawal et al. (2020) evaluate predictive ability when predicting "moral machine" choices, akin to the trolley problem (Foot, 1967).

Fudenberg et al. (2022) and Fudenberg et al. (2023) respectively develop the measures of completeness and restrictiveness, which we adopt here to evaluate a model's prediction accuracy and flexibility. Fudenberg et al. (2022) calculate the completeness of models predicting certainty equivalents for binary lotteries under risk (as well as predicting initial play in matrix games and human generation of random sequences). Fudenberg et al. (2022) 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. (2023) show that CPT achieves much higher completeness then a two-parameter specification generated by Disappointment Aversion, proposed by Gul (1991), but CPT is also substantially less restrictive.

Ke and Zhao (2023) axiomatize locally linear utility functions. Combined with results from Arora et al. (2018), they show that multilayer perceptron neural networks with rectified linear unit activation functions are equivalent to continuous finite piecewise linear functions, which in turn are characterized by completeness, transitivity, continuity, and a weakening of independence. In that sense, our exercise is partially an analysis comparing high dimension parametric utility functions and low dimensional ones. Hsieh et al. (2023) generalize the structure from Ke and Zhao (2023) to describe a logit neural network. They fit their model to estimate the frequency of choosing binary alternatives in a lottery and show that their model outperforms both EUT and CPT. They also find that EUT outperforms CPT in out-of-sample prediction.

We additionally contribute to an expansive experimental literature analyzing models of choice under risk and mention similar recent work below. Fudenberg and Puri (2022) utilize a similar framework to evaluate the completeness of EUT and all combinations of (cumulative) prospect theory (PT, CPT) and simplicity theory at the aggregate and group levels, and they use machine learning as the estimate of irreducible error for completeness purposes. In contrast, we assume perfect prediction at the individual level and evaluate how well economic models perform against machine learning models. Peterson et al. (2021) use neural networks as a parametric utility function (generally capable of a wide search space) to compare the frameworks of EUT, PT, and CPT at the aggregate level. Overall, they find more support for PT than for CPT, especially as data size increases. Clithero et al. (2023) compare the performance of the Becker et al. (1964) mechanism and machine learning models at making predictions of willingness to pay (WTP) for snack foods in a laboratory setting. They find that inputting stated WTP from a BDM mechanism from multiple subjects into a machine learning model to predict purchase decisions at a given price provides a better fit over the BDM prediction itself. The results are robust to using binary choice data instead of WTP, despite binary choice data only containing ordinal information, and using both types of data to improve machine learning models further.

In a traditional vein of experimental economics, Bernheim and Sprenger (2020) utilize a three-good experiment to conduct a nonparametric test of CPT that exploits implied variation in $\mathrm{MRS}_{\mathrm{YZ}}$ when a third variable $X$ passes from $X > Y$ to $X < Y$. They find very low variation, unable to reject rank *independence*. Dembo et al. (2021), utilizing one of the data sets analyzed below, use measures of consistency with utility maximization, monotonic utility maximization, and expected utility maximization to evaluate whether deviations from EUT are caused by the independence axiom, or more core axioms of monotonicity, completeness, and/or transitivity. They find that "[v]iolations of EUT ... run deeper than violations of independence, challenging the most prominent non-EUT alternatives." In a similar vein, Halevy et al. (2023) find that, conditional on violating subjective expected utility in a choice under ambiguity experiment, subjects also violate extremely general versions of non-SEU models such as smooth ambiguity averse preferences or variational preferences.

We share the view of Peysakhovich and Naecker (2017) that individual heterogeneity requires behavior to be examined at an individual level. 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.

# 3 Template for Analysis

## 3.1 Experimental Design

Our dataset is comprised of data from laboratory experiments in which subjects solve a portfolio choice problem. The data was collected by Dembo et al. (2021) to study choice under risk and similar experimental data was collected by Ahn et al. (2014) to study choice under ambiguity. To differentiate the experiments from the choice domain, we capitalize Risk and Ambiguity when discussing the experiments and their associated data. The experiments were conducted at the University of California, Berkeley and the University of California, Los Angeles. The 168 subjects in the Risk experiment and the 154 subjects in the Ambiguity experiment were recruited from all undergraduate classes at these institutions. The experimental procedures described below are identical to those described by Choi et al. (2007c) and used by Choi et al. (2007a) except the present design involves three states (instead of two) and three associated Arrow securities.

More precisely, there are assumed to be three equally probable states of nature, $s = 1, 2, 3$, and an Arrow security for each state. Let $S$ denote the set of states. An Arrow security for state $s$ is defined to be a promise to deliver one dollar if state $s$ occurs and nothing otherwise. Let $\mathbf{x} = (x_1, x_2, x_3) \geq \mathbf{0}$ denote a portfolio of securities, where $x_s$ denotes the number of units of security $s$. A portfolio $\mathbf{x}$ must satisfy the budget constraint $\mathbf{p} \cdot \mathbf{x} = 1$, where $\mathbf{p} = (p_1, p_2, p_3) \geq \mathbf{0}$ is the vector of security prices and $p_s$ denotes the price of security $s$.

Each experimental session consisted of 50 independent rounds of decision problems. These decision problems were presented using a graphical interface. On a computer screen, subjects saw a graphical representation of a three-dimensional budget set $\mathcal{B} = \{\mathbf{x} \mid \mathbf{p} \cdot \mathbf{x} = 1\}$. Each of the axes corresponds to one of three accounts, $x$, $y$ and $z$. The subject's decision problem is to choose a portfolio from the budget set; that is, to allocate his wealth among the three accounts while satisfying the budget constraint. The budget sets selected for each subject in different rounds were independent of each other and of the sets selected for any of the other subjects in their rounds.[3]

At the beginning of each round, the experimental program dialog window went blank and the entire setup reappeared. Subjects could use the mouse or the keyboard arrows to move the pointer on the computer screen to the desired portfolio. Choices were restricted to portfolios on the budget set, so that subjects could not violate a balanced budget. The process was repeated until all 50 rounds were completed. Subjects were told that the payoff in each round was determined by the number of tokens in each account and that, at the end of each round, the computer would randomly select one of the accounts, $x$, $y$ or $z$, for the payoff.

In the Risk experiment, subject were informed that each account, $x$, $y$ or $z$, was *equally likely* to be selected. In the Ambiguity experiment, subjects were only informed that one of the accounts was selected with probability $\frac{1}{3}$ whereas the other two accounts were selected with *unknown* probabilities that sum to $\frac{2}{3}$. In practice, the probability of one of the "ambiguous" states was drawn from the uniform distribution over $[0, \frac{2}{3}]$. This distribution was not announced to the subjects. If the distribution had been revealed to the subjects, the decision problem would have involved compound risk rather than ambiguity. The account with the known probability was held constant throughout a given experimental session but its labeling, $x$, $y$ or $z$, was changed across sessions.

During the course of the experiment, subjects were not provided with any information about the account that had been selected in each round. Instead, at the end of the experiment, the experimental program randomly selected one round from each participant and used the subject's choice from that round's decision problem to determine the subject's payoff. Each round had an equal probability of being chosen, and the subject was paid the amount they had earned in that round. Note that by selecting a single round for the payoff, we prevent subjects from diversifying their risk across the 50 rounds. See Ahn et al. (2014) and Dembo et al. (2021) for an extended description of the experimental design and procedures, as well as full experimental instructions that include screenshots of the computer program dialog windows.

## 3.2 Evaluating Model Performance

The evaluation is twofold, following Fudenberg et al. (2023). First, models are evaluated on their out-of-sample prediction performance. We use the completeness measure from Fudenberg et al. (2022), which reports the fraction of predictable variation a model captures. Second, models are evaluated on their flexibility. We use the restrictiveness measure from Fudenberg et al. (2023), which reports the relative distance of a model to synthetic data. A more complete model better captures the regularities in the data, but the model might have enough flexibility to accommodate any

[3] The axes were scaled from 0 to 100 tokens and are held constant throughout a given experimental session. For each round, the computer randomly selected a budget set subject to the constraints that each intercept lies between 0 and 100 tokens and at least one intercept must be greater than 50 tokens. Payoffs were calculated in terms of tokens and then converted into dollars, where each token was worth $0.50.

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, as discussed in Section 4.

We relate these two measures with a subject-level measure of consistency with rationality. The most basic question to ask about choice data is whether it is consistent with individual utility maximization. In principle, the presence of three states could cause not just a departure from EUT (under risk) and SEU (under ambiguity), but a more fundamental departure from rationality. Thus, before calibrating and evaluating particular utility functions, we first test whether choices can be utility-generated. Afriat (1967) and Varian (1982, 1983) show that choices from a finite number of budget sets are consistent with maximization of a well-behaved (piecewise linear, continuous, increasing, and concave) utility function if and only if they satisfy the Generalized Axiom of Revealed Preference (GARP). Since GARP offers an exact test (either the data satisfy GARP or they do not), we assess how nearly individual choice behavior complies with GARP by using the Critical Cost Efficiency Index (CCEI) of Afriat (1972), 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.[4]

We evaluate model performance at the individual-level, and thus suppress subject indicators in the definitions below. Otherwise, we introduce notation relevant for the analysis. Given a chosen (i.e. "demanded") portfolio $\mathbf{x}$, det $d_s$ be the (relative) demand of state $s$, defined as $d_s = \frac{x_s}{\sum_{s'} x_{s'}}$. Note that since budget sets are required to satisfy $\mathbf{p} \cdot \mathbf{x} = 1$ with equality, knowing the relative demand for two states is sufficient to know the chosen portfolio. We denote the vector of relative demands as $\mathbf{d} = (d_1, d_2)$. In all analysis below, model performance is evaluated over $\mathbf{d}$-space instead of $\mathbf{x}$-space.

Let $\boldsymbol{B}$ denote the set of budget lines, and let $\{(\mathcal{B}^i, \mathbf{d}^i)\}_{i=1}^{50}$ denote the data observed for an individual. Following the terminology and notation of Fudenberg et al. (2023), a *predictive mapping* $f : \boldsymbol{B} \to \mathbf{d}$ is a map from budget sets into relative demand. Mappings are evaluated using the mean-squared error (MSE) *loss function* $\ell : \mathbf{d} \times \mathbf{d} \to \mathbb{R}$ where $\ell(f(\mathcal{B}^i), \mathbf{d}^i) = ||f(\mathcal{B}^i) - \mathbf{d}^i||^2$ is the error

[4] Bronars (1987) builds on Becker (1962) and compares the behavior of our actual subjects to the behavior of simulated subjects who randomize uniformly on each budget line. The power of the Bronars (1987) test is defined to be the probability that a random subject violates GARP. The broad range of budget sets faced by each subject provides a rigorous test of GARP. In particular, the changes in endowments and relative prices are such that budget lines cross frequently. As a result, all the random subjects have violations, implying the Bronars criterion attains its maximum value. See Dembo et al. (2021) for more details on the power tests of revealed preference conditions.

assigned to a predicted relative demand $f(\mathcal{B}^i)$ when the actual relative demand is $\mathbf{d}^i$.[5] The expected prediction error for a mapping $f$ is the expected loss

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

where $P$ denotes the joint distribution of $(\mathcal{B}, \mathbf{d})$.[6] We are interested in comparing set of mappings parametrized by some $\Theta$, $\mathcal{F}_\Theta = \{f_\theta\}_{\theta \in \Theta}$, which we call "models" or "parametric models". The prediction error of a parametric model $\mathcal{F}_\Theta$ is denoted by the lowest expected prediction error of mappings contained in that model:

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

where $f^*_\Theta = \arg\min_{f \in \mathcal{F}_\Theta} \mathcal{E}_P(f)$. Models that share an overarching theme are referred to as model "families".

### 3.2.1 Completeness

Completeness is the amount that a model 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 model $\mathcal{F}_\Theta$, denoted by $\kappa_\Theta$, 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 ideal mapping with irreducible error is defined by

$$f^*(\mathcal{B}^i) = \arg\min_{\mathbf{d}^*} \mathbb{E}_P[\ell(\mathbf{d}^*, \mathbf{d}^i)|\mathcal{B}^i].$$

Since in either experiment 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 $f^*(\mathcal{B}^i) = \mathbf{d}^i$, which in turn implies that $\mathcal{E}_P(f^*) = 0$.

The naive baseline mapping $f_n$ is assumed to be i.i.d uniform choice over $\mathbf{d}$. Recall that when given a subject's actual choice $\mathbf{d}$, the (squared) error of a naive mapping is:

$$\ell(\mathbf{d}_n, \mathbf{d}) = ||\mathbf{d}_n - \mathbf{d}||^2 = (d_1 - d_{1,n})^2 + (d_2 - d_{2,n})^2$$

The uniform random naive mapping assumes that the relative demand is drawn uniformly over the

[5] Recall that the $p$-norm is $||\mathbf{x}||_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$. When a norm is written without a $p$ subscript, we mean the 2-norm.

[6] Note that $P$ for a model of deterministic demand would have a degenerate conditional distribution for $\mathbf{d}$ when $\mathcal{B}$ is known. Additionally, because we conduct analysis at the individual level, $P$ may be different for each subject.

region $\Omega = \{(d_1, d_2) \mid d_1, d_2 \geq 0 \text{ and } d_1 + d_2 \leq 1\}$. The probability distribution function of this distribution is

$$f(d_1, d_2) = \begin{cases} 2 & (d_1, d_2) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

For a given demand $\mathbf{d} = (d_1, d_2)$, the expected prediction error of a random draw is:

$$\begin{aligned} &\int\int_{\Omega} \left[(d_1 - \eta_1)^2 + (d_2 - \eta_2)^2\right] f(\eta_1, \eta_2) d\eta_1 d\eta_2 \\ &= 2\int_0^1 \int_0^{1-\eta_2} \left[(d_1 - \eta_1)^2 + (d_2 - \eta_2)^2\right] d\eta_1 d\eta_2 \\ &= \frac{1}{3} - \frac{2}{3}d_1 - \frac{2}{3}d_2 + d_1^2 + d_2^2 \end{aligned}$$

We estimate $\mathcal{E}_P(f_n)$ and $\mathcal{E}_P(f^*_\Theta)$ for each subject, and then plug these estimates into the formula for $\kappa_\Theta$ for an estimate for completeness. The estimate of $\mathcal{E}_P(f_n)$ is $\hat{\mathcal{E}}_n = \frac{1}{50}\sum_{i=1}^{50} \frac{1}{3} - \frac{2}{3}d_1^i - \frac{2}{3}d_2^i + (d_1^i)^2 + (d_2^i)^2$. To estimate the expected prediction error of parametric models, $\mathcal{E}_P(f^*_\Theta)$, we use 10-fold cross-validation. In this exercise, the set of data $\{(\mathcal{B}^i, \mathbf{d}^i)\}_{i=1}^{50}$ is partitioned into 10 equally sized, mutually exclusive subsets $Z_1, \ldots, Z_{10}$, each with five observations. Each partition $Z_k$ is then used for out-of-sample prediction, where the complement of the partition $Z_{-k}$ is used to estimate $f^*_\Theta$ as $\hat{f}^{-k} = \arg\min_{f_\theta \in \mathcal{F}_\Theta} \frac{1}{45}\sum_{i \notin Z_k} \ell(f_\theta(\mathcal{B}^i), \mathbf{d}^i)$. The estimate $\hat{f}^{-k}$ is then used to generate an estimated out-of-sample prediction error over $Z_k$, $\hat{e}_k = \frac{1}{5}\sum_{i \in Z_k} \ell(\hat{f}^{-k}(\mathcal{B}^i), \mathbf{d}^i)$. The estimate of $\mathcal{E}_P(f^*_\Theta)$, denoted $\hat{\mathcal{E}}_\Theta$, is the average of the partition-level error estimates:

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

The estimate of completeness is thus

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

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

### 3.2.2 Restrictiveness

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

We use the space of representative agents, bootstrapped from actual data, as our distribution of permissible mappings $\mu$. Real subject choices from the Risk experiment are pooled together and divided into 10 equally sized bins based on the price ratio between the cheapest Arrow security and the second cheapest Arrow security. Within each bin, again responses are divided into 10 equally sized bins based on the price ratio between the second cheapest Arrow security and the most expensive Arrow security. For each observed budget set, a relative demand is drawn from a uniform distribution over the choices from that budget set's bin. We group the budget sets by subject, resulting in 168 "subjects" with synthetic data.

We estimate restrictiveness by separately estimating the numerator and the denominator. Like completeness, the estimate of the uniformly random naive baseline mapping requires no parameter estimation, and can be immediately evaluated on the given data. We estimate the numerator of restrictiveness by evaluating *within-sample* error across the distribution of "subjects" generated, or as $\frac{1}{168}\sum_{j=1}^{168}\frac{1}{50}\sum_{i=1}^{50}\ell(f_\theta^j(\mathcal{B}^{i,j}), f^j(\mathcal{B}^{i,j}))$, where $f^j$ is the $j^{\text{th}}$ "subject", $f_\theta^j = \text{argmin}_{f \in \mathcal{F}_\Theta} d(f, f^j)$, and $\mathcal{B}^{i,j}$ is the $i^{\text{th}}$ budget set of the $j^{\text{th}}$ "subject". Fudenberg et al. (2023) show that the asymptotic difference between this restrictiveness estimate and the actual restrictiveness value is normal with mean zero.

## 3.3 Economic Models

### 3.3.1 The "Generalized Kinked" Specification

There is a variety of theoretical models of attitudes toward risk and ambiguity. We skip the development and analysis of the models and refer the interested reader to Ahn et al. (2014) for more details. The main specification of Ahn et al. (2014) – the so-called generalized kinked specification – builds on Quiggin's (1982) Rank-Dependent Utility (RDU). This three-parameter (parsimonious) specification provides measures of attitudes towards risk, loss and ambiguity. To illustrate such preferences, first suppose there are three states and that the probabilities of all states are objectively known and equiprobable. This corresponds to the Risk experiment, with $\pi_s = \frac{1}{3}$ for all $s$. The RDU of the rank-ordered portfolio $\tilde{\mathbf{x}}$ takes the form

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

where $\beta_L, \beta_M, \beta_H > 0$ are decision weights that sum to unity, $\tilde{\mathbf{x}} = (x_L, x_M, x_H)$ is a *rank-ordered* portfolio with payoffs $x_L \leq x_M \leq x_H$, and $u$ is the Bernoulli index. Note that this formulation reduces to Expected Utility Theory (EUT) when $\beta_L = \beta_M = \beta_H$ (since each state has an equal likelihood of occurring) and encompasses a number of prominent non-EUT models.

In the RDU model, a weighting function $w : [0,1] \rightarrow [0,1]$ transforms probabilities into probability weights. The weighting function is assumed to be increasing and satisfies $w(0) = 0$ and $w(1) = 1$. With pure risk, the probability weight of each payoff depends only on its (known) probability and its ranking and can be expressed in terms of $w$ as follows:

$$\begin{aligned} &\beta_L = w\left(\tfrac{1}{3}\right), \\ &\beta_M = w\left(\tfrac{2}{3}\right) - w\left(\tfrac{1}{3}\right), \\ &\beta_H = 1 - w\left(\tfrac{2}{3}\right). \end{aligned}$$

The decision weights assigns to each payoff the probability weight of receiving at most that payoff, less the probability weight of receiving strictly less than that payoff. The model assumes zero probability assigned to anything less than $x_L$ and probability one assigned to anything more than $x_H$.

In order to include ambiguity, we assume that state 2 has an objectively known probability $\pi_2 = \frac{1}{3}$, whereas states 1 and 3 occur with unknown probabilities $\pi_1$ and $\pi_3$, and that the unknown probabilities $\pi_1$ and $\pi_3$ are skewed using the weights $0 \geq \alpha \geq 1$ and $1 - \alpha$ for the low $x_{\min} = \min\{x_1, x_3\}$ and high $x_{\max} = \max\{x_1, x_3\}$ payoffs, respectively. The decision weight of each payoff now depends on whether its probability is unknown. The utility takes the form

$$U(x_L, x_M, x_H) = \begin{cases} w_1 u(x_2) + (w_2 - w_1) u(x_{\min}) + (1 - w_2) u(x_{\max}) & x_2 \leq x_{\min} \\ w_3 u(x_{\min}) + (w_2 - w_3) u(x_2) + (1 - w_2) u(x_{\max}) & x_{\min} \leq x_2 \leq x_{\max} \\ w_3 u(x_{\min}) + (w_4 - w_3) u(x_{\max}) + (1 - w_4) u(x_2) & x_{\max} \leq x_2, \end{cases}$$

and its ranking and can be expressed in terms of the weighting function $w$ as follows:

$$\begin{aligned} w_1 &= w\left(\tfrac{1}{3}\right), \\ w_2 &= w\left(\tfrac{2}{3}\alpha + \tfrac{1}{3}\right), \\ w_3 &= w\left(\tfrac{2}{3}\alpha\right), \\ w_4 &= w\left(\tfrac{2}{3}\right). \end{aligned}$$

Ahn et al. (2014) adopt a simpler three-parameter version of this model, in which the parameter $\delta$ measures the attitudes towards ambiguity and the parameter $\gamma$ measures attitudes towards loss. The mapping from the two parameters $\delta$ and $\gamma$ to the four parameters $w_1, ..., w_4$ is given by the equations

$$\begin{aligned} w_1 &= \tfrac{1}{3} + \gamma, \\ w_2 &= \tfrac{2}{3} + \gamma + \delta, \\ w_3 &= \tfrac{1}{3} + \gamma + \delta, \\ w_4 &= \tfrac{2}{3} + \gamma. \end{aligned}$$

where $-\frac{1}{3} \leq \delta, \gamma \leq \frac{1}{3}$ and $-\frac{1}{3} \leq \delta + \gamma \leq \frac{1}{3}$. When $\delta = 0$ , we obtain the (ambiguity neutral) RDU model.[7] When $\gamma = 0$, we obtain the (loss neutral) $\alpha$-Maxmin Expected Utility ($\alpha$-MEU) model of Ghirardato et al. (2004) and Olszewski (2007).[8] If $\gamma > 0$ (preferences are loss averse), there is a nondifferentiable kink at all portfolios where $x_s = x_{s'}$, for any $s$ and $s'$. If $\delta > 0$ (preferences are also ambiguity averse) the kink at portfolios where $x_1 = x_3$ is sharper than the kink at portfolios where $x_1 = x_2$ and $x_2 = x_3$. Through a suitable change of variables, we can also interpret the case where $\gamma, \delta > 0$ as reflecting Recursive Nonexpected Utility (RNEU) proposed by Segal (1987, 1990). When $\delta = 0$ and $\gamma = 0$, we have the standard Subjective Expected Utility (SEU) representation.

As in Ahn et al. (2014), we assume that risk preferences are represented by a utility function

[7] In the disappointment aversion model of Gul (1991) the indifference curves have a kink only at the safe portfolio $x_1 = x_2 = x_3$, whereas in the generalized kinked specification the indifference curves have a kink at all portfolios where $x_s = x_{s'}$ for some $s \neq s'$.

[8] It can also be derived as a special case of a variety of other models in the literature, including Maxmin Expected Utility (Gilboa and Schmeidler, 1989), Choquet Expected Utility (Schmeidler, 1989), and Contraction Expected Utility (Gajdos et al., 2008). See Ahn et al. (2014) more precise details.

$u(x)$ with constant absolute risk aversion (CARA), $u(x) = -e^{-\rho x}$, where $x$ is the number of tokens and $\rho$ is the coefficient of absolute risk aversion. This specification has two advantages. First, it is independent of the (unobservable) initial wealth level of the subjects. Second, it accommodates portfolios where $x_s = 0$ for some state $s$ even when initial income is zero. For each subject and for each specification, we generate estimates of the risk ($\rho$), loss ($\gamma$) and ambiguity ($\delta$) parameters using nonlinear least squares (NLLS). Like Ahn et al. (2014), in the econometric analysis, we do not restrict the parameters so that preferences are always not loss loving or ambiguity loving, but we do restrict the risk parameter so that preferences are always risk averse ($\rho \geq 0$).[9]

### 3.3.2 The smooth specification

An alternative prominent utility specification that captures attitudes towards ambiguity is a model that is differentiable everywhere. In this model, the agent has a subjective (second-order) distribution over the possible (first-order) prior beliefs over states. Unsure which of the possible first-order prior beliefs actually governs the states, the agent transforms the expected utilities for all prior beliefs by some concave function before integrating these utilities with respect to her second-order distribution. This procedure is entirely analogous to the transformation of wealth into cardinal utility before computing expected utility under risk. The concavity of this transformation captures ambiguity aversion. We refer to this model as Recursive Expected Utility (REU), owing to its recursive double expectation.[10]

The general form of the REU model is

$$U(\mathbf{x}) = \int_{\Delta S} \varphi \left( \int_S u(x_s)\, d\pi(s) \right) d\psi(\pi),$$

where $\psi \in \Delta(\Delta(S))$ is a (second-order) distribution over possible priors $\pi$ on the state space $S$ and $\varphi : u(\mathbb{R}_+) \to \mathbb{R}$ is a possibly nonlinear transformation over expected utility levels. Here, $\Delta(\Delta(S))$ denotes the space of all probability measures over $\Delta(S)$, the set of all probability distributions on $S$. We reduce the REU model to two parameters by assuming that $\varphi(z) = -\frac{1}{\alpha} e^{-\alpha z}$, which replicates the constant curvature of $u$, and that $\psi$ is uniformly distributed over the set of priors consistent with the objective information $\pi_2 = \frac{1}{3}$ and $\pi_1 + \pi_3 = \frac{2}{3}$. This specification is characterized by two parameters: one is the familiar coefficient of risk aversion and the other is a measure of attitudes towards ambiguity.

The estimation of the smooth specification proved more difficult than the generalized kinked

[9] In the absence of ambiguity, risk-seeking individuals always allocate all their tokens to the cheapest Arrow security. This is also the behavior that would be implied by risk neutrality so the attitude toward risk is immaterial and hence cannot be estimated. In the presence of ambiguity aversion, the implications of risk-seeking behavior are not quite so stark, but the difficulty of identifying the underlying risk preferences remains.

[10] See the models of Ergin and Gul (2009), Klibanoff et al. (2005), Nau (2006), Giraud (2014), Halevy and Feltkamp (2005), Ahn (2008), and Seo (2009), among others.

specification. Adding a non-convexity to the utility function would add to the computational problems. We thus follow Ahn et al. (2014) and restrict the parameters so that preferences are always ambiguity (and risk) neutral/averse. An additional problem with the smooth specification is the difficulty of interpreting the parameters, which are not truly identified without some auxiliary assumption about cardinal utility. As a result, Ahn et al. (2014) do not discuss the estimated parameters of the smooth specification in detail.

## 3.4 Machine Learning Models

We consider three main families of machine learning models – regularized regressions, tree-based, and neural networks. Each family is commonly used in machine learning, and increasingly economics, literature. We include multiple approaches because there is no declared 'winning' method.[11] For each subject, we consider the most complete of all models considered.[12] For each model, we encode the budget set $\mathcal{B}^i$ as a three dimensional vector of the intercept $1/p_1$ and price ratios $p_2/p_1$ and $p_3/p_1$. We abuse notation by referring to both the budget set, as well as the vector representing the budget set, as $\mathcal{B}^i$. For the Ambiguity experiment, we permute subjects' choices such that the second state is assigned to always be the state with known probability $\frac{1}{3}$.

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

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

where $\hat{\beta}$ solves

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

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

[12] 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.

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

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

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

The parameter $\lambda$ affects the degree to which the size of $\beta$ affects the objective function. If $\lambda = 0$, then the optimization is OLS. We use leave-one-out cross-validation to determine the parameter $\lambda \in [0, 0.2, 0.4, 0.6, 0.8, 1]$. The parameter vector $\theta$ for regularized regressions models is a linear coefficient vector.

**Tree-based** Let $t$ denote one of the possible variables associated with a budget set. Unlike the linear relationship assumed in regularized regression, tree-based models divide the set of budget sets $\boldsymbol{B}$ into subsets (based on the prices and the endowment) and estimate a model on each of the subsets. This division is done recursively. That is, given some index of observations $Z$ corresponding to data $\{(\mathcal{B}^i, \mathbf{d}^i)\}_{i \in Z}$, the the algorithm considers all further binary partitions that can be represented as separating data based on a variable $x$ being above or below a given threshold $k$: $\{(\mathcal{B}^i, \mathbf{d}^i)\}_{i \in Z \text{ and } t^i \leq k}$ and $\{(\mathcal{B}^i, \mathbf{d}^i)\}_{i \in Z \text{ and } t^i > k}$. Of these partitions, the selected partition is the $(t, k)$ pair that minimizes error when applying optimal models to each partition.

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

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

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

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

The standard decision tree model, denoted Mean, takes the sample mean token share $x$ of each subset. We use Mean as well as two extensions. The first extension, known more broadly as model trees (Quinlan and Others (1992)), changes the estimated model from a sample mean to a linear regression (Linear). Mean is nested in Linear. The last tree-based model, the random forest model (RF) averages the decision rules of multiple standard decision trees. Each tree is given a bootstrapped data set, and is generally seen as an improvement over singular decision trees (Breiman (2001)). In addition to limits on depth and minimum sample size, RF regulates the number of trees, which we choose to be 10, 50, and 100 trees. Because each tree is 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 $\theta$.

**Neural networks** A Neural network, specifically a multilayer perceptron, transforms budget sets into relative demand predictions by nonlinear regression, whose functional form assumes a series of nested transformations. In our setup, the transformation takes two parts. Recall that a budget set $\mathcal{B}$ is encoded as a three dimensional vector containing $1/p_1, p_2/p_1$, and $p_3/p_1$. First, this vector undergoes an affine transformation $W^{(0)}\mathcal{B} + b^{(0)}$, where $W^{(0)}$ and $b^{(0)}$ are a weight matrix and bias vector of size $n_0 \times 3$ and $n_0 \times 1$, respectively. The dimension $n_0$ is prespecified by the analyst, and the exact weights and biases are estimated from data. Second, the affine transformation is again transformed by a function $\sigma^{(0)} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_0}$ to obtain a new vector $\mathcal{B}^{(1)} = \sigma^{(0)}(W^{(0)}\mathcal{B} + b^{(0)})$. The function $\sigma$ is also prespecified by the analyst. The resulting vector, $\mathcal{B}^{(1)}$, is referred to as a "hidden layer". It is then used as the input to generate another hidden layer, $\mathcal{B}^{(2)} = \sigma^{(1)}(W^{(1)}\mathcal{B} + b^{(1)})$, using a new affine transformation defined by $\underset{n_1 \times n_0}{W^{(1)}}$ and $\underset{n_1 \times 1}{b^{(1)}}$ as well as transformation by $\sigma^{(1)}$. This process continues for the number of hidden layers prespecified by the analyst. The final affine transformation results in a scalar value that can be interpreted as the estimated relative demand.

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

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

We use the L-BFGS algorithm (Liu and Nocedal, 1989) to estimate $W^{(i)}$ and $b^{(i)}$. This is a standard optimization algorithm that uses first and second order information to iteratively update estimates of parameter values. This algorithm is generally not feasible to compute for larger models and larger data sets, but is applicable in our setting with 50 observations per subject.[13] The estimation objective function to be minimized is mean squared error, which is the same objective function used to evaluate all models (through completeness and restrictiveness). For example, consider a network of 2 hidden layers each with dimension 15. The corresponding objective function is:

$$\min_{\substack{W^{(0)}_{15\times 2}, W^{(1)}_{15\times 15}, W^{(2)}_{1\times 15}, b^{(0)}_{15\times 1}, b^{(1)}_{15\times 1}, b^{(2)}_{1\times 1}}} \sum_{x^i \in \mathcal{D}} \ell(f(W,b), \mathbf{d}^i) = || \mathbf{d}^i - W^{(2)} \sigma \left( W^{(1)} \sigma \left( W^{(0)} \mathcal{B}^i + b^{(0)} \right) + b^{(1)} \right) - b^{(2)} ||^2$$

# 4 Power Analysis

In this section, we investigate whether completeness scores from our experimental data can capture variation between different types of data generating processes (DGPs). We examine three classes of DGPs of choice under risk, and vary parameters within each class.

1. First, we consider a non-EUT (RDU) DGP, varying the level of probability weighting $\gamma$. Increasing levels of $\gamma$ should reduce the completeness of EUT while maintaining the completeness of RDU.

2. Second, we consider a DGP which implements EUT with logistic noise, and we vary the level of noise. Increasing noise should reduce completeness, regardless of model. However, more flexible models such as RDU are able to better fit noise due to their additional parameters, and thus may be prone to overfitting when the true underlying model remains EUT.

3. Finally, we implement two DGPs that violate the assumptions of both EUT and RDU: one that violates monotonicity with respect to first-order stochastic dominance (FOSD), and one that violates both FOSD and GARP.[14] Because both EUT and RDU satisfy GARP and FOSD, if there is a systematic deviation from these axioms, the additional parameter from RDU should not provide any additional out-of-sample prediction over EUT. The rules are simple enough such that machine learning models should have higher completeness than economic models for both DGPs.

[13] (For a full treatment, see Bottou et al. (2018) and Sun et al. (2019)).

[14] We say a mapping $f$ satisfies monotonicity with respect to first-order stochastic dominance if for all $\mathcal{B}$, there is no portfolio $\mathbf{x}'$ such that the portfolio implied by $\mathbf{d} = f(\mathcal{B})$, $\mathbf{x} = \frac{1}{p_1 d_1 + p_2 d_2 + p_3(1-d_1-d_2)}(d_1, d_2, 1-d_1-d_2)$ has the property that $\min\{x'_1, x'_2, x'_3\} \geq \min\{x_1, x_2, x_3\}$, $\text{median}\{x'_1, x'_2, x'_3\} \geq \text{median}\{x_1, x_2, x_3\}$, and $\max\{x'_1, x'_2, x'_3\} \geq \max\{x_1, x_2, x_3\}$. Almost all models that extend EUT also satisfy this feature, including all non-EUT and non-SEU models in Section 3.3.2. For shorthand, we say $f$ "satisfies FOSD".

For each DGP, we compare the completeness of both EUT and RDU. Each DGP is run on the same 50,000 budget lines randomly generated in the same manner as Section 3.1. The 50,000 budget lines are partitioned in groups of 50 to form 1,000 simulated subjects. Note that since completeness is a data-based measure, we recalculate completeness for each DGP.

## 4.1 Non-EUT (RDU) Subjects

We first examine a non-EUT (RDU) DGP with linear Bernoulli utility and varying degrees of probability weighting. We use the generalized kinked framework from Section 3.3.2, setting $\delta = 0$. Note that linear utility is not a special case of CARA utility which we use to estimate demand. Hence, there is some misspecification of our parametric RDU estimations as well. We use four specifications of the RDU parameter $\gamma \in \{0, 1/100, 1/20, 1/12\}$. Recall that $\gamma = 0$ is EUT, and increasing $\gamma > 0$ displays greater levels of pessimism, distorting probability from the cheapest Arrow security to the most expensive Arrow security. Thus $\gamma$ translates to decision weights $\beta_\gamma = (\beta_L, \beta_M, \beta_H) = (1/3 + \gamma, 1/3, 1/3 - \gamma)$ for the lowest, middle, and highest accounts respectively.

Figure 1 shows a heatmap of completeness scores for EUT and RDU for each specification of $\gamma$. The completeness of EUT is plotted on the x-axis, and the completeness of RDU is plotted on the y-axis. Above and to the right of the heatmap are the marginal completeness distributions approximated using kernel density functions with a Gaussian kernel.

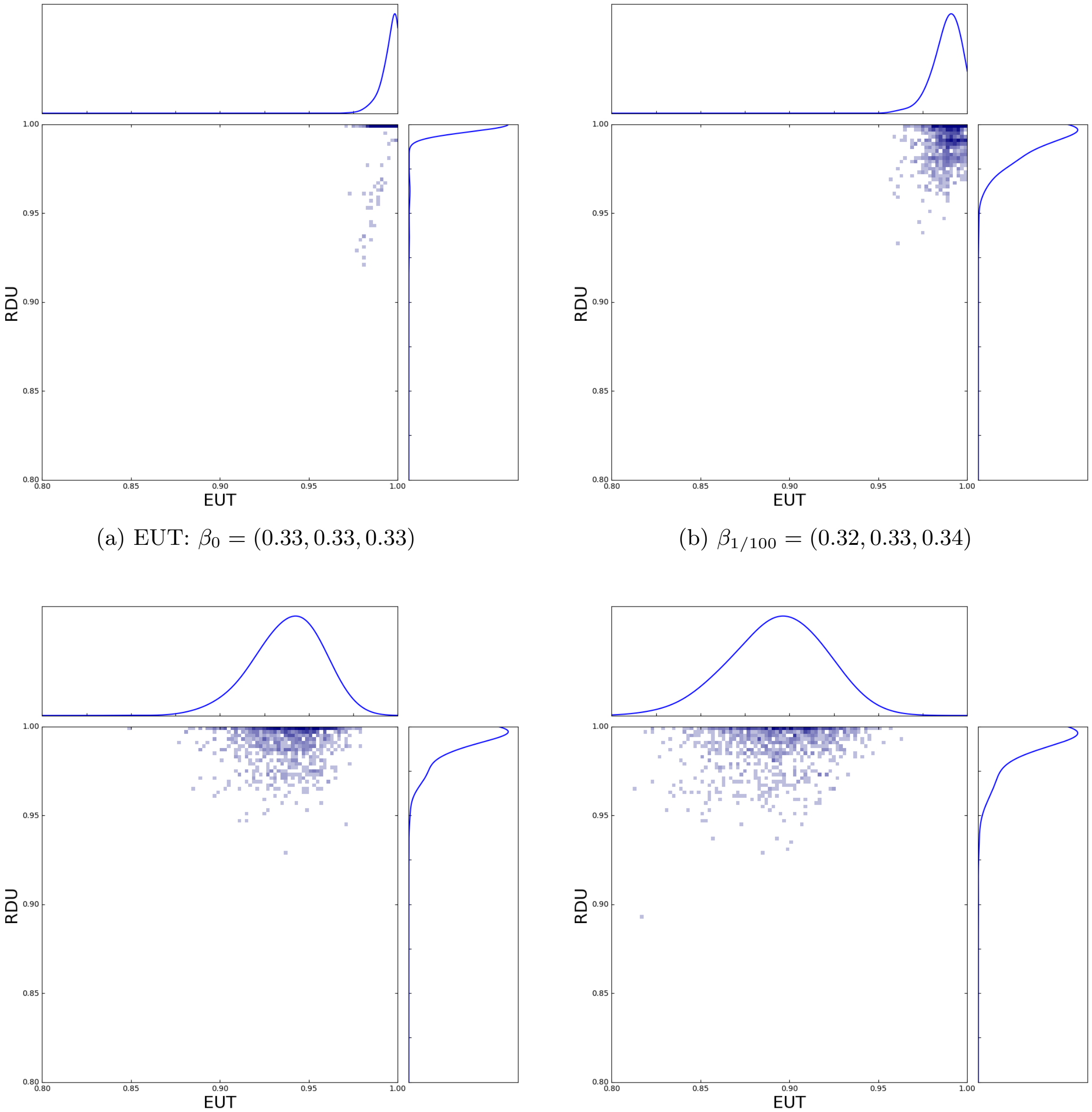
We observe a marked downward shift in completeness of EUT as probability distortion increases, with no equivalent downward shift for RDU. Completeness for EUT in Panel 1a is almost always 100%: note the extremely high concentration of subjects in the top-right corner of the heatmap. A marginal change from $\beta_0$ to $\beta_{1/100}$, which corresponds to a one percentage point probability distortion, moves the distribution leftward, but with a minimal effect on EUT completeness: the average EUT completeness is 98.9%. With subsequent increases in distortion, the average EUT completeness continues to decrease until 89.2% at $\beta_{1/12}$. Increasing distortion further does not seem to negatively affect EUT completeness: in Figure A.2, we show figures for $\beta_{1/9}$, $\beta_{1/6}$, and $\beta_{1/3}$, where the average EUT completeness are 87.5%, 87.9%, and 86.9% respectively. Note that $\beta_{1/3}$ is the maximal possible well defined distortion. On the contrary, the completeness of RDU remains at 99% or higher for all $\beta_\gamma$.

## 4.2 Noisy EUT Subjects

Next, we examine DGPs with EUT preferences that make choices according to a utility-weighted logistic distribution over the budget plane. We assume a CARA Bernoulli utility with $\rho = 0.065$ as the coefficient of absolute risk aversion.[15] The probability of a specific allocation $\mathbf{x}$ being chosen

[15] The CARA utility parameter value 0.065 was selected by conducting an empirical exercise to determine the CARA utility parameter "closest" to log utility. We chose a CARA utility specification, instead of simply using

Figure 1: The completeness of EUT and RDU on RDU DGPs with various levels of $\gamma$.



(a) EUT: $\beta_0 = (0.33, 0.33, 0.33)$

(b) $\beta_{1/100} = (0.32, 0.33, 0.34)$

(c) $\beta_{1/20} = (0.28, 0.33, 0.38)$

(d) $\beta_{1/12} = (0.25, 0.33, 0.42)$

The completeness of EUT is plotted on the x-axis, and the completeness of RDU is plotted on the y-axis. Subjects are binned in a $100 \times 100$ grid of completeness scores from 80% to 100%. Bin colors range from white to navy, where lighter colors correspond to bins with fewer subjects, and darker colors correspond to more subjects. Above and to the right of the heatmap are the marginal distributions approximated using kernel density functions with a Gaussian kernel.

from a budget set with prices $\mathbf{p}$ is

$$P(\mathbf{x}) = \frac{\exp\left[\nu \mathbb{E}u(\mathbf{x})\right]}{\int\int_{\mathbf{x}'|\mathbf{px}'=1} \exp\left[\nu \mathbb{E}u(\mathbf{x}')\right] dx_1' dx_2'}$$

As $\nu$ approaches zero, the distribution approaches uniform random choice over the budget set. As $\nu$ approaches infinity, the distribution approaches deterministic utility maximization. We simulate four levels of $\nu \in \{0, 0.25, 1, 10\}$.

Figure 2 shows the distribution of completeness scores for RDU and EUT for each DGP. Overall, we observe a strong positive correlation between completeness and $\nu$. As $\nu$ increases, we observe a large monotonic shift towards the upper right corner; in general, the supports of completeness scores for either model barely overlap between different levels of $\nu$. Additionally, we see little separation between the completeness of EUT and RDU: most simulated subjects lie on or around the 45-degree line. The main difference between model performances comes in the variance of completeness differences between EUT and RDU: when $\nu = 0$, the variance between EUT and RDU completeness differences is 0.011; when $\nu = 10$, the variance is 0.00018.

Most importantly, despite being the nested model, the completeness of EUT is higher than that of RDU for at least 70% of subjects. Additionally, as a second order effect, the proportion of subjects for whom EUT has higher completeness increases with $\nu$: EUT has higher completeness than RDU for 70.7% of subjects when $\nu = 0$, and increases monotonically with $\nu$ to 78.2% when $\nu = 10$.

## 4.3 Irrational Subjects

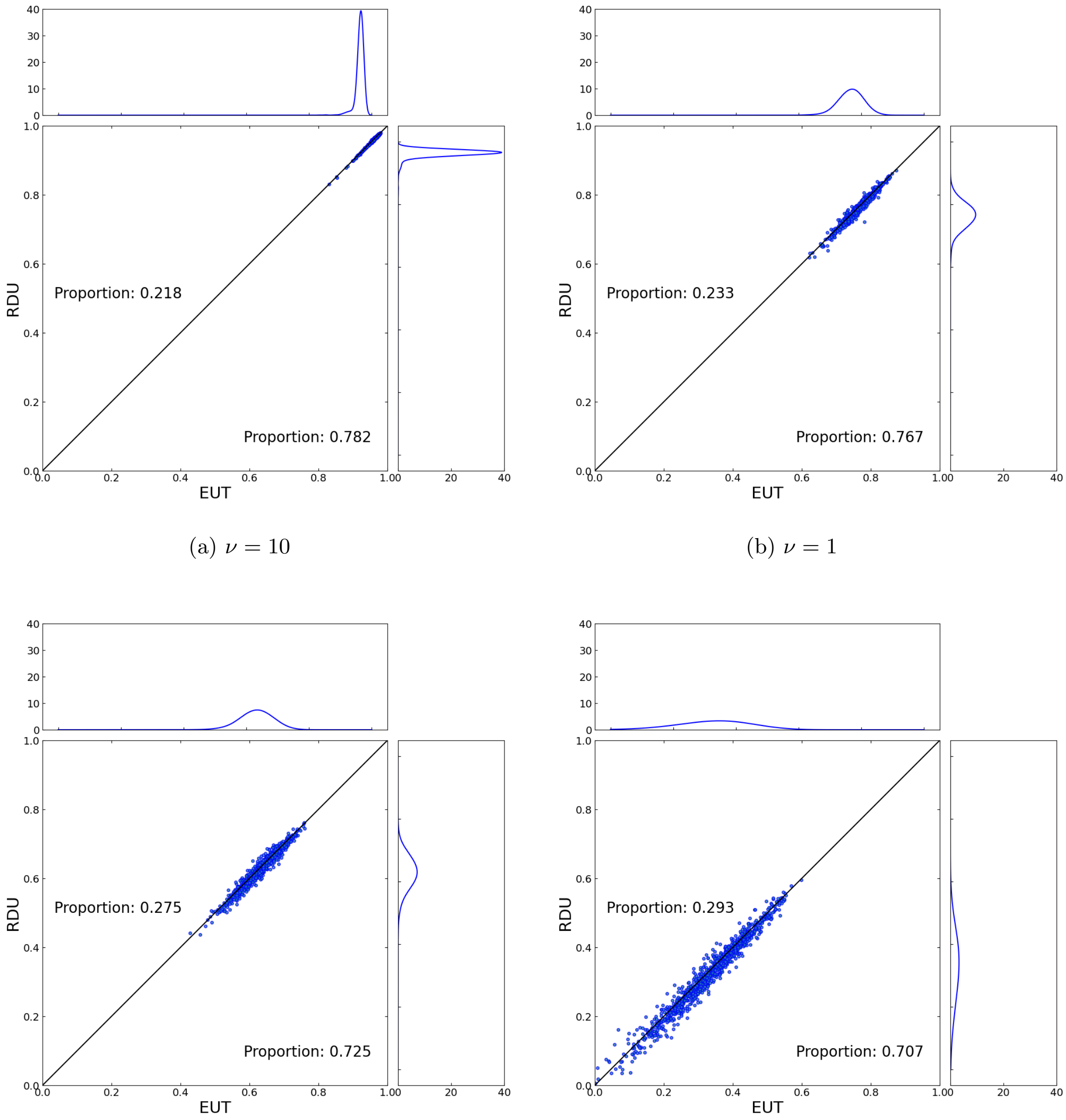
Finally, we consider two DGPs of extremely irrational behavior.[16] First, we consider a DGP that exclusively chooses portfolios that allocates all income to the most expensive Arrow security. In the experiment, this corresponds to exclusively choosing the account with the lowest intercept. This DGP violates both GARP and FOSD. Second, we consider a DGP that exclusively chooses portfolios that allocate all income to an arbitrary account, regardless of price. In the experiment, this corresponds to exclusively choosing the same account in each round. Without loss of generality we assume the arbitrary account is $x$. This DGP almost always violates just FOSD, but always

log utility, to ensure that whatever utility is used to generate the distribution will be a special case of the models estimated.

Using the 500 budget lines associated with the first 10 of the real 168 subjects, we calculated the demand estimates of a log utility maximizer, where the demanded $(x_1, x_2, x_3) = (\frac{1}{3p_1}, \frac{1}{3p_2}, \frac{1}{3p_3})$, equating expenditures. We then searched the space of CARA utility parameters that minimized the mean squared error of the specification's estimates with those of log utility. The resulting value was 0.065.

[16] While "rationality" typically refers to complete and transitive preferences, we believe that FOSD is so fundamental to choice under risk that violating it can be deemed, at least colloquially, irrational.

Figure 2: The completeness of EUT and RDU on noisy EUT DGPs with various levels of $\nu$.



(a) $\nu = 10$

(b) $\nu = 1$

(c) $\nu = 0.25$

(d) $\nu = 0$

In each panel, the central plot shows the completeness of two models for each subject, along with the proportion of subjects for whom the model has a higher completeness. The completeness of EUT is plotted on the x-axis, and the completeness of RDU is plotted on the y-axis. Finally, we provide a kernel density estimation, with a Gaussian kernel, showing the marginal distribution of completeness scores for each model.

satisfies GARP.[17] Since both EUT and RDU assume consistency with GARP and FOSD, these subjects fall outside of the framework of both models.

Figure 3 shows the completeness of EUT and RDU for both DGPs. Panel 3a displays the completeness of subjects who always chooses the lowest intercept. The performance of RDU and EUT are nearly identical for each subject, and thus all subjects fall within a small margin of the 45-degree line. Panel 3b displays the completeness of subjects who always choose $x_1$. In this case, RDU and EUT performance are no longer perfectly correlated, but still exhibit strong positive correlation in completeness. This is in line with theoretical predictions discussed in Appendix B: the "lowest intercept" DGP generates theoretical guarantees for identical behavior between EUT and RDU regardless of the exact budget lines asked, whereas this is not the case for the "always $x$" DGP.

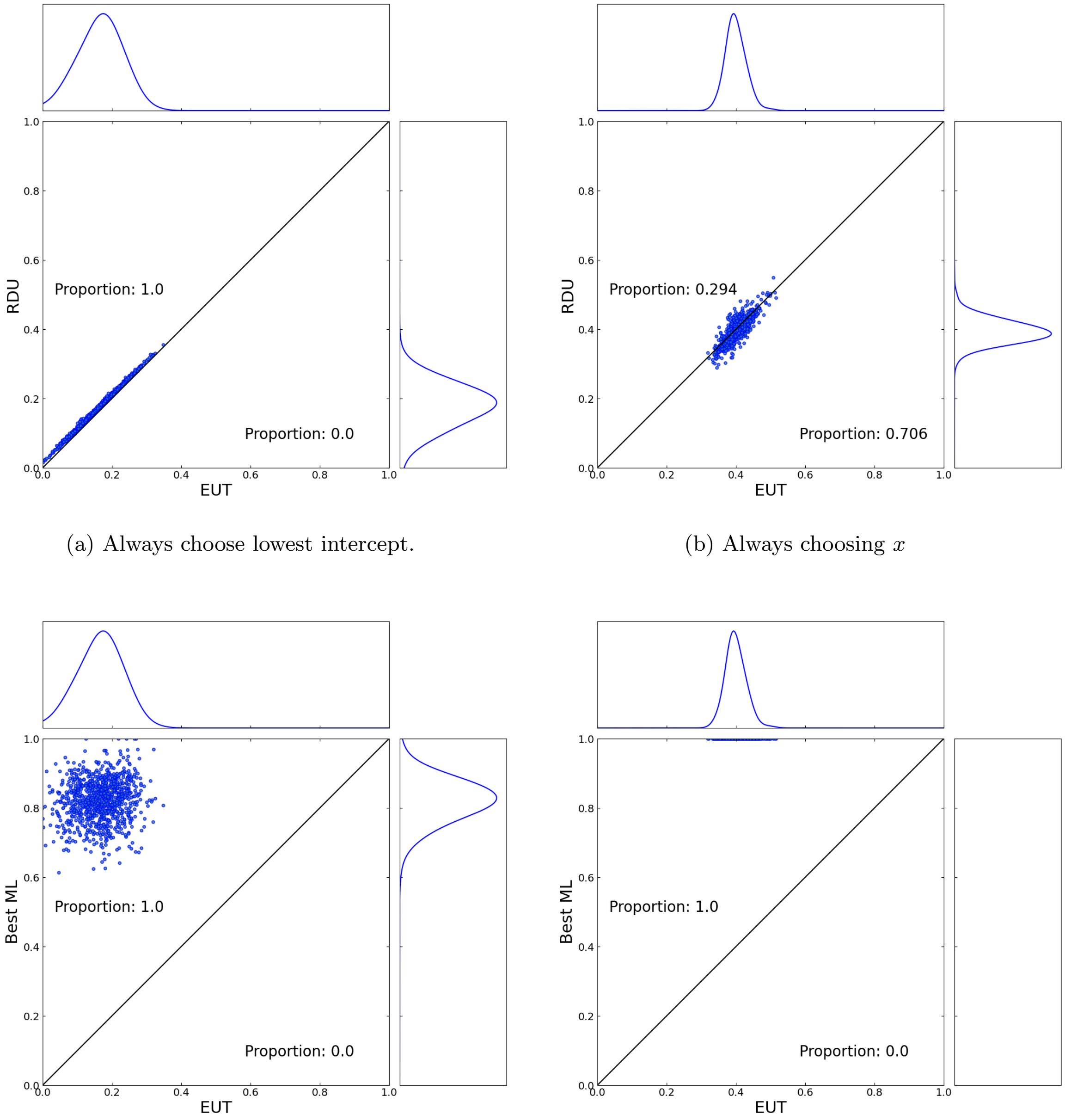
Also, while the results show little difference between EUT and RDU in both cases, there is an increase in completeness when moving from Panel 3a, where both GARP and FOSD are violated, to Panel 3b, where only FOSD is violated. Hence, in this case, satisfying additional axioms of both EUT and RDU increases completeness for both models.

Panels 3c and 3d replicate the analysis, instead comparing the performance of the most complete machine learning model per subject, denoted "Best ML", to that of EUT in both scenarios. We use a subset of models discussed in Section 3.4 for computational purposes, omitting linear model trees. Panel 3c shows that machine learning models, which do not make assumptions of GARP or FOSD, are not fully complete when subjects always choose the lowest intercept. However, Best ML is significantly more complete than EUT, as denoted by the concentration of subjects in the upper-left corner of the diagram far from the diagonal. The inputs we use for the machine learning model have an impact. For example, note that a simple OLS would be able to perfectly predict if we include a variable indicating when an intercept is the lowest. In contrast, Panel 3d shows that machine learning models are fully complete. This result comes naturally from the fact that demand is constant, regardless of variation in input variables.

In summary, we establish that the experiment is sufficiently powerful to detect true differences between economic models. Within our framework, we confirm the degree to which RDU completeness differs from EUT completeness on such a data set, and that the difference can be made meaningfully large even when minutely modifying the RDU parameter away from zero. Second, when the underlying DGP is "EUT plus noise", EUT has higher completeness for the vast majority of subjects, even with low levels of noise. This is especially notable because, in theory, nested models always have lower completeness than their nesting model. However, the out-of-sample estimate of completeness explicitly removes this guarantee, because the additional flexibility of the nesting

[17] In the case where subjects always choose a single security, FOSD is not violated if that security always has the highest intercept (lowest price).

Figure 3: Scatterplot of completeness of EUT and RDU on simulated “pathological cases” data.



(a) Always choose lowest intercept.

(b) Always choosing $x$

(c) Always choose lowest intercept.

(d) Always choosing $x$

In each panel, the central plot shows the completeness of two models for each subject, along with the proportion of subjects for whom the model has a higher completeness. The completeness of EUT is plotted on the x-axis, and the completeness of RDU is plotted on the y-axis. Finally, we provide a kernel density estimation, with a Gaussian kernel, showing the marginal distribution of completeness scores for each model.

model may be able to fit noise in the data. We show that the latter effect can be stronger in our analysis. Finally, we show that EUT and RDU have nearly identical performance when analyzing a DGP that violates FOSD but not GARP, and when analyzing a DGP that violates both. Hence, our experiment is powerful enough to distinguish between models even when the data does not the limit regarding size. In addition, the this separation does not occur because the assumptions of both models are violated.

# 5 Results

In this section, we examine the results of the analysis between economic and machine learning models. We first examine broad results across choice domains, and then examine the relative performance of models within choice domains.

Table 1 provides a population-level summary of our results, which we refer to throughout the analysis. All economic models and machine learning models are present, together with a report of results for the most complete machine learning model, denoted "ML". Panel A reports results in the Risk experiment, and Panel B reports results in the Ambiguity experiment. Panel C reports the results from Ellis et al. (2022), who conduct a similar analysis on subjects making decisions in an analogous two-dimensional risky budgetary choice environment (two equiprobable states, two Arrow securities, etc). In each panel, the first row reports the average completeness of each model across all subjects. The next two rows report the average completeness of each model for subjects below and above median CCEI, respectively. The final row reports the restrictiveness of the model. For regularized regressions, tree-based models, and Best ML, we report restrictiveness as weighted averages of the most complete model in the family for each subject.

There are two main insights from comparing choice domains. First, when comparing Panel A to Panel C, we show that adding a dimension slightly reduces completeness, but greatly increases restrictiveness of all models. The economic model restrictiveness increase is larger than that of machine learning models in both relative and absolute terms. Second, all models exhibit a slight increase in completeness from Risk to Ambiguity. However, our results are consistent with CCEI scores between Risk and Ambiguity subjects: the average CCEI's are 93.8% and 94.5% for Risk and Ambiguity, respectively.

**Panel A: Risk**

| Completeness | EUT/SEU | RDU | $\alpha$-MEU | Smooth | RNEU | Reg | Trees | NN | ML | $\Delta_{EUT-ML}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| All subjects | 85.1 | 85.2 | - | - | - | 74.6 | 81.4 | 68.7 | 81.9 | 3.1 |
| Below median CCEI | 79.7 | 79.5 | - | - | - | 70.7 | 75.5 | 62.7 | 76.3 | 3.4 |
| Above median CCEI | 90.5 | 90.9 | - | - | - | 78.6 | 87.4 | 74.6 | 87.6 | 2.9 |
| Restrictiveness | 48.1 | 47.5 | - | - | - | 31.6 | 17.1 | 28 | 20.3 | - |

**Panel B: Ambiguity**

| Completeness | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| All subjects | 85.4 | - | 85.3 | 85.5 | 84.9 | 78.3 | 82.5 | 72.4 | 83.1 | 2.3 |
| Below median CCEI | 80.3 | - | 80.3 | 80.5 | 79.6 | 74.2 | 77.3 | 67 | 78.2 | 2.2 |
| Above median CCEI | 90.5 | - | 90.3 | 90.5 | 90.2 | 82.4 | 87.6 | 77.8 | 88 | 2.5 |
| Restrictiveness | 48.1 | - | 47.9 | 47.5 | 47.6 | 31.6 | 17.1 | 28 | 20.3 | - |

**Panel C: Risk 2D**

| Completeness | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| All subjects | 89.3 | 89.2 | - | - | - | 79.2 | 89.1 | 71.6 | 89.6 | -0.3 |
| Below median CCEI | 82.6 | 82.3 | - | - | - | 75.4 | 82.5 | 69.5 | 83.3 | -0.7 |
| Above median CCEI | 95.9 | 96 | - | - | - | 83 | 95.8 | 73.6 | 95.8 | 0.1 |
| Restrictiveness | 18.6 | 16.6 | - | - | - | 20.7 | 9.4 | 14.4 | 11.4 | - |

Table 1: Completeness and restrictiveness of models

## 5.1 Economic Models

Next, we compare completeness and restrictiveness scores between economic models. Figure 4 plots the completeness of the classical models of EUT and SEU against non-EUT models: RDU (Panel 4a), $\alpha$-MEU (Panel 4b), Smooth Ambiguity (Panel 4c), and RNEU (Panel 4d). For each panel, the central plot shows the completeness of both models for each subject, along with the proportion of subjects for whom the model has a higher completeness.

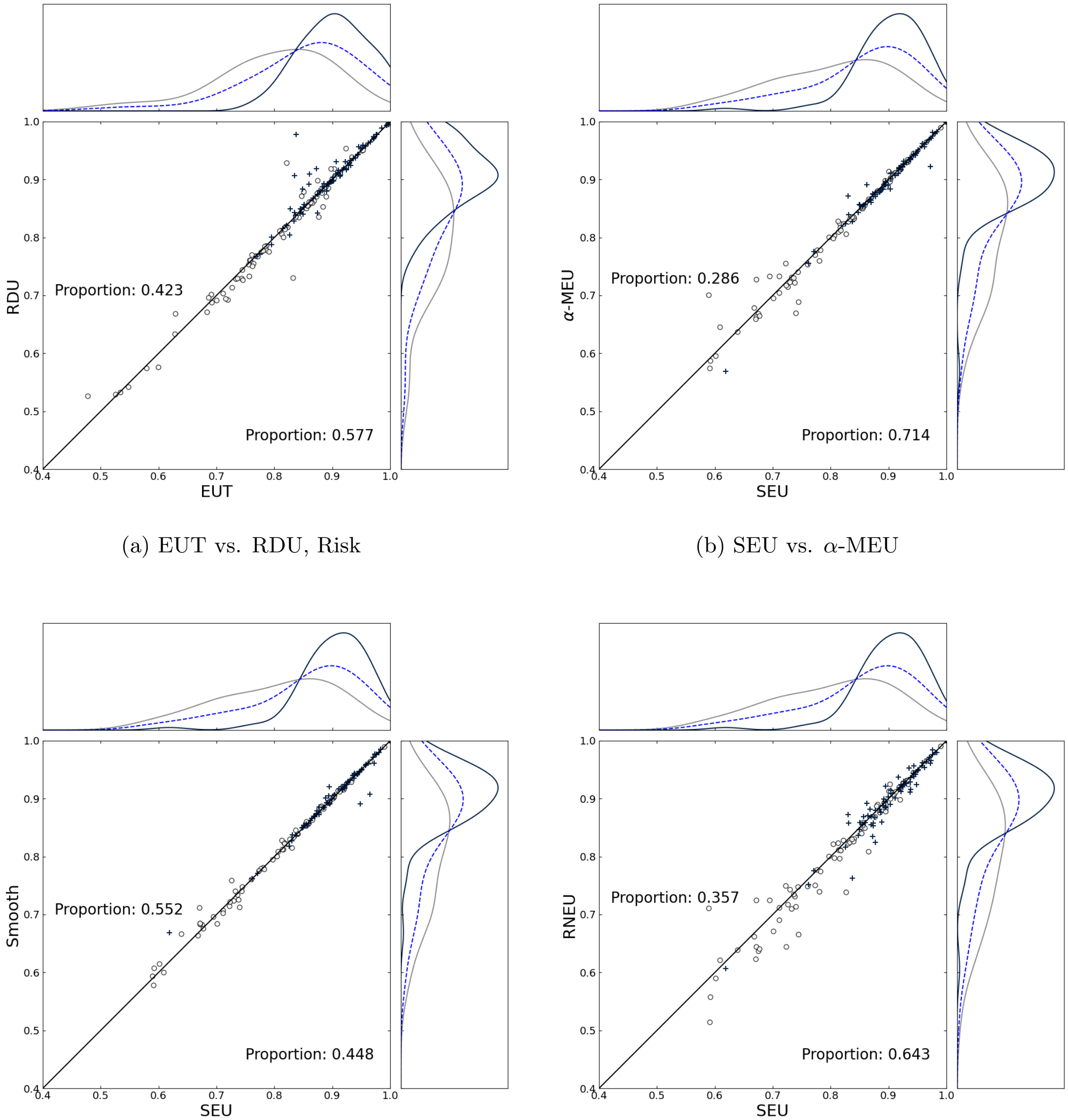
**Risk** First, we examine economic model performance in the Risk experiment, shown in Panel 4a. Overall, EUT and RDU have extremely similar performance metrics. The average completeness of EUT (85.1%) is one-tenth of a percentage point lower than the average completeness of RDU (85.2%). The restrictiveness of EUT (48.1%) is marginally higher than the restrictiveness of RDU (47.5%). Both models exhibit large differences in completeness depending on CCEI. For EUT (RDU), the completeness of subjects below median CCEI is 79.7% (79.5%), whereas for above median CCEI it is 90.5% (90.9%). However, given a subpopulation, there are minor difference between models.

At the individual-level, there is striking symmetry between EUT and RDU completeness. Most subjects are concentrated around the 45-degree line, where the completeness of both models are the same. Additionally, EUT has a higher completeness for 57.7% of subjects. Thus, the distribution of completeness scores are remarkably similar. A Kolmogorov-Smirnov test supports this finding by failing to reject the null hypothesis that the distributions are the same ($p = 0.97$). Next, subjects with above median CCEI, are located more towards the top-right corner, indicating that they have higher completeness for both models. The completeness distribution of subjects scoring above median CCEI subjects, for both EUT and RDU, also has a higher mean and lower variance than that of subjects scoring below median CCEI.

**Ambiguity** Next, we examine economic model performance in the Ambiguity experiment. Here, results are extremely similar to those in the Risk experiment: there are no major differences in economic model performance when examining completeness across all subjects. The average completeness of SEU is 85.4%, which is within half a percentage point of all other models: $\alpha$-MEU (85.3%), Smooth Ambiguity (85.5%), and RNEU (84.9%). Again, there is a large difference in average completeness between subjects scoring below median CCEI and above median CCEI subjects. Given a subpopulation, however the average completeness between models is approximately the same. Additionally, the restrictiveness of each model is within one percentage point.

Similar to the comparison between EUT and RDU in the Risk experiment, Panel 4b shows there is symmetry between completeness scores, with the majority of subjects located near the 45-degree line. Like the Risk experiment, subjects with above-median CCEI exhibit higher completeness for both SEU and $\alpha$-MEU. Additionally, SEU has a higher completeness for 71.4% of subjects,

Figure 4: Scatterplot of completeness of EUT or SEU models compared to non-EUT models.



(a) EUT vs. RDU, Risk

(b) SEU vs. $\alpha$-MEU

(c) SEU vs. Smooth Ambiguity

(d) SEU vs. RNEU

Subjects with below median CCEI are plotted as circles with a gray outline, and subjects with above median CCEI are plotted as navy "+" symbols. The completeness of EUT or SEU is plotted on the x-axis, and the completeness of the non-EUT model is plotted on the y-axis. Finally, we provide a kernel density estimate showing the marginal distribution of completeness scores for each model. The light gray, navy, and dotted blue curves plot the marginal distributions for below median CCEI subjects, above median CCEI subjects, and all subjects, respectively.

implying that the additional parameter of $\alpha$-MEU does not measurably increase out-of-sample predictive performance. The results of Panels 4c and 4d are similar to those between SEU and $\alpha$-MEU. In general, the average completeness of all models are nearly identical. Similar to Risk, the distribution of completeness scores are not jointly significant from each other using an Anderson-Darling $k$-sample test ($p > 0.25$). In Appendix Figure A.1, we show similar figures comparing $\alpha$-MEU to Smooth Ambiguity and RNEU. We find similar results to Figure 4, with model completeness distributed tightly around the 45-degree line, and subjects scoring above median CCEI exhibiting higher completeness.

## 5.2 Economic Models vs. Machine Learning Models

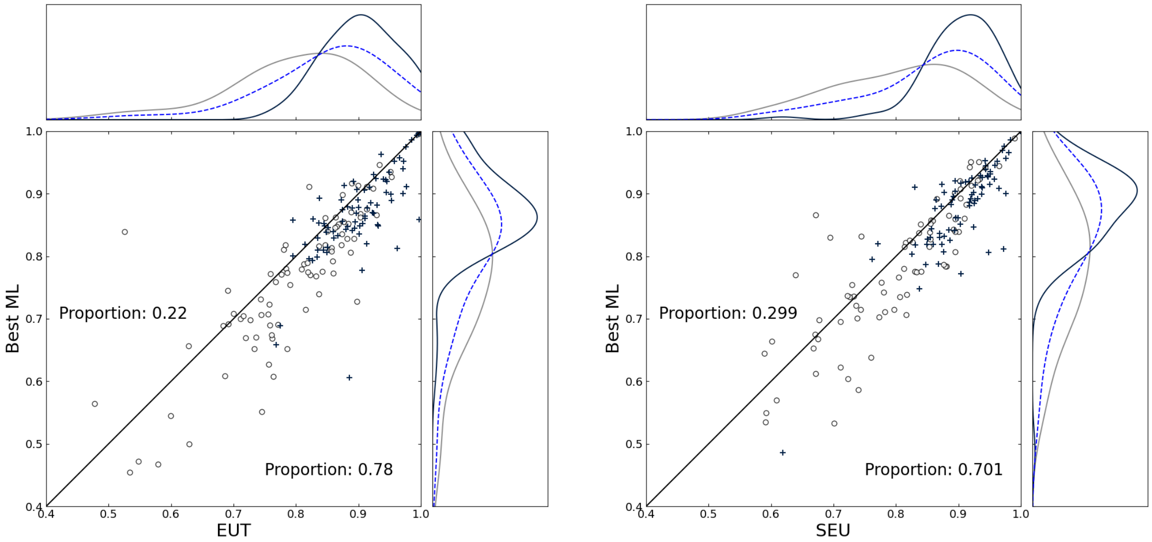
Finally, we examine machine learning model performance. We first report results between machine learning models. Overall, there is a negative correlation between model completeness and model restrictiveness. In the Risk experiment, the most complete machine learning model family, tree-based models (achieving 81.4% of the feasible reduction in prediction error), is the least restrictive model with a restrictiveness of 17.1%. Compared to tree-based models, regularized regressions and neural networks exhibit lower completeness (74.6% and 68.7% respectively) but higher restrictiveness (31.6% and 28.0% respectively). The results are extremely similar in the Ambiguity experiment.

Next, we examine machine learning model performance relative to EUT and SEU. Figure 5 plots the distribution of completeness for EUT or SEU and the most complete machine learning model ("Best ML") in the Risk and Ambiguity experiments. EUT and SEU outperform the most complete machine learning model in Risk and Ambiguity respectively, in both absolute and relative terms. Under Risk, EUT is more complete than the most complete machine learning model for 78.0% of subjects; for SEU in the Ambiguity experiment, it is 70.1%. The completeness distributions are not symmetric (Kolmogorov-Smirnov test, $p = 0.01$), and subjects are less concentrated around the 45-degree line. In the Risk experiment, the most complete machine learning model has a completeness of 81.9%, compared to a completeness of 85.1% for EUT. These results are not driven by differences in model completeness on subjects above or below median CCEI: the most complete machine learning model achieves 3.4 percentage points lower completeness for below median CCEI subjects, and 3.3 percentage points lower completeness for above median CCEI subjects.

In summary, we show that there are slight decreases in completeness when adding a dimension of choice, but large increases in restrictiveness. Hence, the assumptions underpinning economic models more strongly restrict demand in three dimensions than two. Despite this restriction, the relative drop in completeness for Best Econ (4.1 percentage points) is lower than that of Best ML (7.7 percentage points), indicating that the assumptions are the "right" restrictions on feasible choice behavior.

Next, we show that the standard models of EUT and SEU are as complete as all non-EUT

Figure 5: Scatterplots of EUT and SEU against the most complete machine learning model.



(a) Risk

(b) Ambiguity

Subjects with below median CCEI are plotted as circles with a gray outline, and subjects with above median CCEI are plotted as navy "+" symbols. The completeness of EUT or SEU is plotted on the x-axis, and the completeness of the non-EUT model is plotted on the y-axis. Finally, we provide a kernel density estimate showing the marginal distribution of completeness scores for each model. The light gray, navy, and dotted blue curves plot the marginal distributions for below median CCEI subjects, above median CCEI subjects, and all subjects, respectively.

economic models in the Risk and Ambiguity experiments, respectively. Notably, we also observe that EUT and SEU exhibit higher completeness for at least 57% of subjects than non-EUT models, all of which nest EUT or SEU as a special case. While, in theory, completeness should be higher for the nesting model, the out-of-sample estimate of completeness explicitly removes this guarantee. As seen in Section 4, these results are consistent with a DGP of an EUT-weighted logistic distribution over budgets, but inconsistent with a DGP of RDU with parametric misspecification.

In addition to higher completeness, EUT and SEU are also not significantly more restrictive. Hence, the greater model flexibility provided by the additional parameters of non-EUT models is relatively small, and does not impact predictive accuracy. These results are consistent with the notion that violations of EUT do not manifest themselves as violations of the independence axiom, which primarily distinguishes EUT and SEU from non-EUT models, but instead come in the form of violations of more core axioms such as GARP.

Finally, EUT and SEU are shown to be more complete and more restrictive than all machine learning models considered. These results are robust across subjects of different levels of consistency with GARP. Hence, we fail to find regular choice patterns that EUT and SEU cannot incorporate but a machine learning model can. Combined, these two observations confirm that EUT and SEU well capture the behavior of subjects.

## 5.3 Individual-level vs Aggregate-level

We contrast the results from the individual-level analysis with that of an aggregate-level analysis. In this, we analyze the cost of making predictions by assuming a representative agent, or a small number of representative agents, in lieu of individual-level estimation. Note that we provide models with all subject choices, but we do not provide the models with the identity of the subject associated with each choice.[18] Thus, models can only vary demand based on prices, and not on subject identity. This increases the amount of data available to the model, at the cost of imprecision. For economic models, we interpret this exercise as a misspecification analysis, as choice data aggregated from multiple preferences is inevitably misspecified because there is no utility function that pooled choices maximize due to Afriat's Theorem (Afriat, 1967).

We pool subjects in three ways. First, we estimate one model per family on the full data set.[19] Next, we classify subjects as "Above median CCEI" or "Below median CCEI", and estimate one model for each pool of subjects. Finally, we calculate restrictiveness by estimating one model for the entire bootstrapped dataset discussed in Section 3.

Figure 6 plots the completeness of individual-level and aggregate-level analysis. Panels 6a and

[18] See Section 6 for an analysis where we provide all subjects with identifiers.

[19] We modify the tree-based model parameter search space to 2, 6, 10, 14, or 18 for tree depth and minimum samples per terminal leaf, modified from 2, 4, 6, and 8 used for tree depth and 2, 4, 6, 8, and 10 for minimum samples per terminal leaf used to estimate individual-level parameters discussed in Section 3.

6b show models analyzed on all Risk and Ambiguity subjects, respectively. Panels 6c and 6d show the completeness of below median CCEI subjects for Risk and Ambiguity. Finally, Panels 6e and 6f show the same for above median CCEI subjects.

First, when analyzing all subjects, all models except neural networks exhibit decreases in completeness. The completeness reduction is largest for economic models, which exhibit completeness reductions of 9 to 12 percentage points. Note that the reductions are slightly higher for nested models, and thus completeness respects nesting: the completeness of RDU is 1 percentage point above EUT, and the completeness of all Ambiguity models is 76%. In contrast, the most complete machine learning model family, tree-based models, exhibits a completeness reduction of just five percentage points. Neural networks increase in completeness, indicating that misspecification cost does not outweigh the value of increased data.

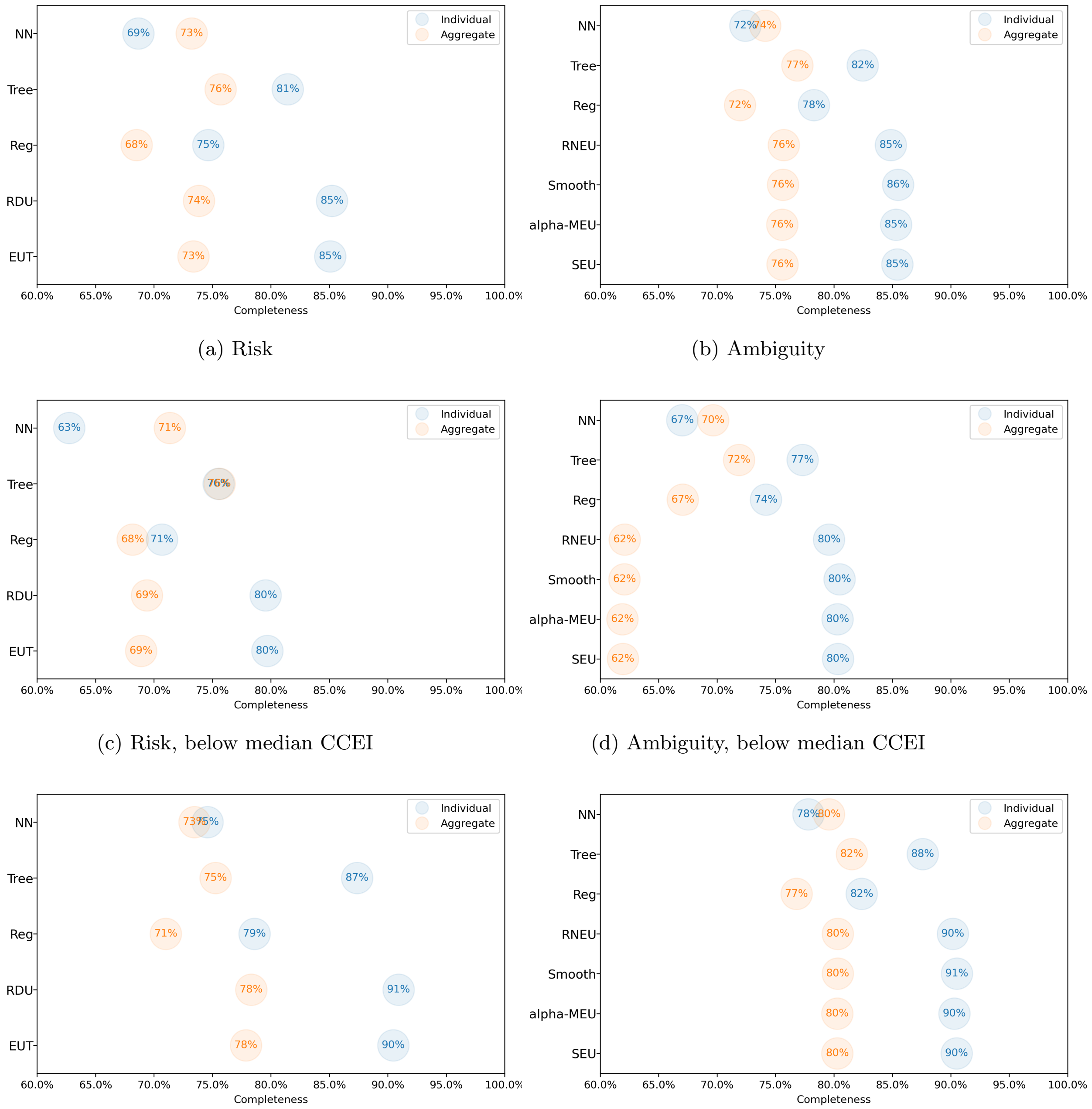
Next, for below median CCEI subjects, aggregate models perform relatively better in Risk than in Ambiguity. For each model, the difference in completeness from individual to aggregate is greater in Risk than Ambiguity. The change is most notable for economic models, which exhibit a drop in completeness of 11 percentage points each under Risk, similar to when analyzing all subjects. A larger drop of 18 percentage points each was found under Ambiguity. Finally, for subjects scoring above median CCEI, we see trends similar to all subjects, but with completeness shifted approximately 5 percentage points higher.

There is asymmetric misspecification cost, and it is largest for economic models. The reduction in completeness is highest for economic models and lowest (or resulted in a completeness increase) for machine learning models. Analyzing results from this pooled data also leads to different conclusions about model quality: the most complete machine learning model is more complete than the most complete economic model. This is in contrast to individual-level analysis, where economic models were more complete in both Risk and Ambiguity. Hence, individual-level economic models are most complete, but machine learning models have the lowest misspecification cost when individual-level data is not available.

## 6 Grouping Subjects

We examine the performance of machine learning models when they are provided with all subject data, including identifying information for each observation. By doing so, we remove misspecification bias and seek to quantify the gains from additional observations from “similar” distributions. We denote this the “Group”, or “Grouped”, analysis. Including this analysis for machine learning models is trivial, and the number of variables can easily grow into the hundreds or thousands. In contrast, aggregating even two subjects’ choices will almost always guarantee a misspecification of economic models due to Afriat’s Theorem.

In particular, for both Risk and Ambiguity, we provide machine learning models with $1/p_1$,



(a) Risk

(b) Ambiguity

(c) Risk, below median CCEI

(d) Ambiguity, below median CCEI

(e) Risk, above median CCEI

(f) Ambiguity, above median CCEI

Figure 6: Change in completeness between individual-level and aggregate-level analysis for models in the Risk experiment

$p_2/p_1$, and $p_3/p_1$ in the same manner as before.[20] Additionally, we provide indicator variables for each subject, as well as an interaction between the subject ID indicator variables and the subject-specific $1/p_1$, $p_2/p_1$, and $p_3/p_1$. When conducting 10-fold cross validation for completeness estimates, we balance subject data by assigning five data points per subject in each fold.

We do not attempt a similar exercise with economic models. In particular, we note that since subjects' choices are independent of other subjects' choices, any joint estimation of preferences, for classical methods such as nonlinear least squares or maximum likelihood estimation, reduces to independent, individual-level estimation. However, it is clear that there exists plausible scenarios where estimating risk preferences on multiple subjects' data as if they were one subject can be beneficial due to the bias-variance tradeoff. Consider a scenario where two subjects implement EUT CARA but with logistic noise, similar to Section 4.2, with different CARA parameter values $\rho_1$ and $\rho_2$. If the noise level is sufficiently high, and if $\rho_1$ and $\rho_2$ are sufficiently close to each other, then estimating one $\hat{\rho}$ for both subjects will be biased but will obtain the benefit of reduced variance because of the larger sample size. Since this approach is sufficiently different, we leave the results for future work.

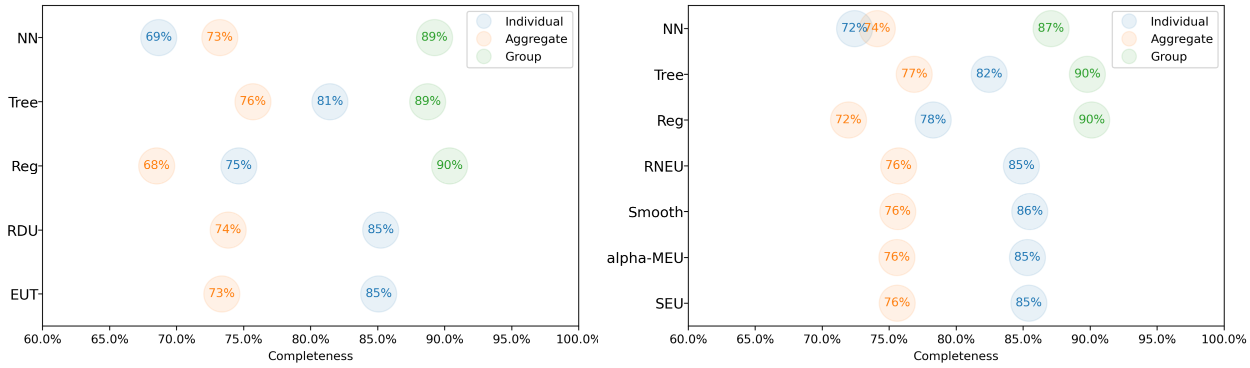
We first show sample-level results in Figure 7. Each model family is presented on the y-axis, and the (average) completeness of individual-level, aggregate-level, and group-level analysis for each model is plotted. There are two findings to note. First, group-level analysis increases completeness for all machine learning models. On average, there is a 13 percentage points completeness increase for a machine learning model when changing from individual-level to group-level analysis. The resulting group-level average completeness scores between machine learning models are within 1 percentage point for Risk, and within 3 percentage points for Ambiguity. Second, and more importantly, all group-level machine learning model families are more complete than all economic models.

We also investigate the per-subject performance of individual-level and group-level modeling. Figure 8 shows the completeness of the most complete group-level machine learning model compared to EUT, SEU and the most complete individual-level machine learning model for Risk and Ambiguity. Across all panels, at least 85% of subjects are better explained by group-level ML than by any other model. Although there is a positive correlation between individual-level completeness and group-level completeness, the relationship is relatively flat. While high completeness subjects have similar completeness regardless of level, low completeness subjects exhibit a significant difference. Specifically, while group-level analysis never results in completeness below 70%, individual-level analysis can yield completeness scores below 50%.

This difference largely affects subjects scoring below median CCEI, who have lower completeness than subjects above median CCEI for individual-level models. While the average completeness is

[20] As discussed in Section 5.3, we again modify the tree-based models' parameter search space to 2, 6, 10, 14, or 18 for tree depth and minimum samples per terminal leaf.

Figure 7: The completeness of models by data structure



(a) Risk

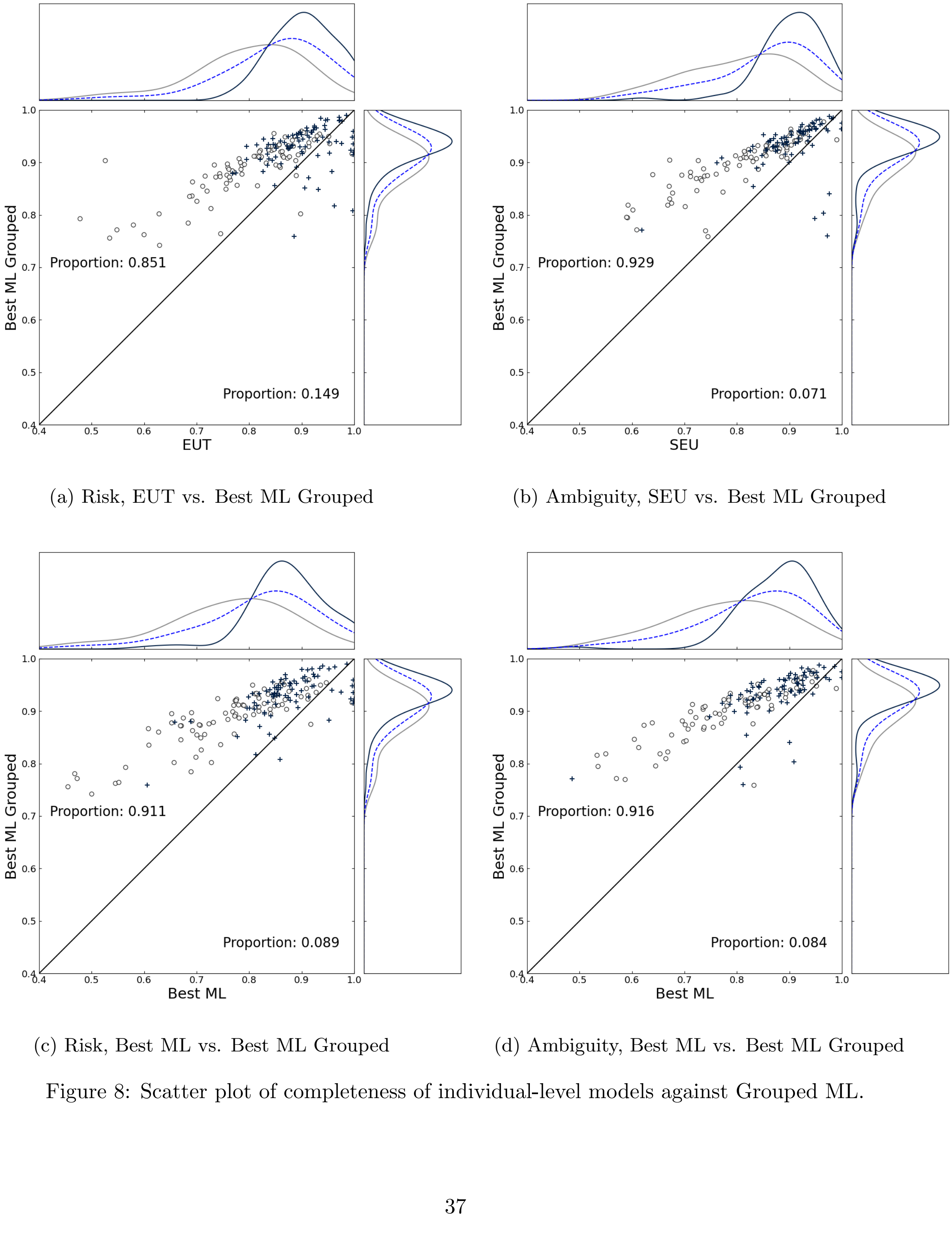
(b) Ambiguity

higher for group-level ML (90.8%) than individual-level EUT (85.1%), the difference is higher for subjects below median CCEI (88.8% compared to 79.7%) than for subjects above median CCEI (92.9% compared to 90.5%). Additionally, subjects for whom individual-level EUT is better than group-level ML overwhelmingly have higher CCEI. For Risk, of the 25 subjects for whom EUT is more complete than group-level ML, 20 have scores above median CCEI.

These results indicate that ML models are able to effectively incorporate trends between subjects. The benefits are largest for subjects below median CCEI, whose choice patterns are less consistent. Group-level ML also is more complete than individual-level EUT and SEU. Thus, the objective of the analyst, as well as the availability of data, determines model selection. In scenarios where data availability is a concern or restraint; or where easy interpretability/explanation of results is necessary, then the analyst should use an economic model. However, in scenarios where easy interpretability is not a concern and when there is ample data, off-the-shelf machine learning techniques outperform economic models.

# 7 Conclusion

We investigate the accuracy of economic models and the strength of their axiomatic foundation using complementary methods of completeness (Fudenberg et al., 2022) and restrictiveness (Fudenberg et al., 2023). We use budgetary choice environments with three dimensions to provide a strong test. Overall, we show that economic models capture individual behavior well, in particular the standard model of expected utility theory. EUT is more complete than non-EUT models, even in choice under ambiguity. This result in consistent with the notion that violations of EUT stem from more fundamental axioms such as GARP instead of violations from the independence axiom.



(a) Risk, EUT vs. Best ML Grouped

(b) Ambiguity, SEU vs. Best ML Grouped

(c) Risk, Best ML vs. Best ML Grouped

(d) Ambiguity, Best ML vs. Best ML Grouped

Figure 8: Scatter plot of completeness of individual-level models against Grouped ML.

Additionally, EUT is more complete than individual-level machine learning models. However, group-level machine learning models, which input information about all subjects into the model, are more complete than economic models, indicating that machine learning models can effectively transfer information across subjects to improve prediction.

Our results bring up important questions about the extent to which additional data can improve predictions, and the transferability of economic preferences. For example, in our group-level analysis, we included data for all subjects, which is a function of the number of subjects. There are multiple ways this can be modified. A naive approach would be to group subjects randomly and input data only from that subject's group. More sophisticated approaches can carefully selet the group, either by using sociodemographic information from surveys such as the Understanding America Survey or the Dutch CentERpanel. Using just choice information, for each subject, one could include data from other subjects only if estimates of EUT risk preferences are "sufficiently close". Additionally, we only investigate "within-domain, across-subject" transferability, but "within-subject, across-domain" is of interest as well. We plan to address these questions in future work.

# References

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

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

Agrawal, M., J. C. Peterson, and T. L. Griffiths (2020): "Scaling Up Psychology via Scientific Regret Minimization," *Proceedings of the National Academy of Sciences*, 117, 8825–8835.

Ahn, D., S. Choi, D. Gale, and S. Kariv (2014): "Estimating Ambiguity Aversion in a Portfolio Choice Experiment," *Quantitative Economics*, 5, 195–223.

Ahn, D. S. (2008): "Ambiguity Without a State Space," *The Review of Economic Studies*, 75, 3–28.

Arora, R., A. Basu, P. Mianjy, and A. Mukherjee (2018): "Understanding Deep Neural Networks With Rectified Linear Units," in *International Conference on Learning Representations*.

Athey, S. and G. W. Imbens (2019): "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, 11, 685–725.

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

Becker, G. S. (1962): "Irrational Behavior and Economic Theory," *Journal of Political Economy*, 70, 1–13.

Bernheim, B. D. and C. Sprenger (2020): "On the Empirical Validity of Cumulative Prospect Theory: Experimental Evidence of Rank-Independent Probability Weighting," *Econometrica*, 88, 1363–1409.

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.

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.

——— (2007c): "Revealing Preferences Graphically: An Old Method Gets a New Tool Kit," *American Economic Review*, 97, 153–158.

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

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

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

Ellis, K., S. Kariv, and E. Ozbay (2022): "What Can the Demand Analyst Learn From Machine Learning?" *Working Paper.*

Ergin, H. and F. Gul (2009): "A Theory of Subjective Compound Lotteries," *Journal of Economic Theory*, 144, 899–929.

Foot, P. (1967): "The Problem of Abortion and the Doctrine of the Double Effect," *Oxford Review*, 5, 5–15.

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

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

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

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

Gajdos, T., T. Hayashi, J.-M. Tallon, and J.-C. Vergnaud (2008): "Attitude Toward Imprecise Information," *Journal of Economic Theory*, 140, 27–65.

Ghirardato, P., F. Maccheroni, and M. Marinacci (2004): "Differentiating Ambiguity and Ambiguity Attitude," *Journal of Economic Theory*, 118, 133–173.

Gilboa, I. and D. Schmeidler (1989): "Maxmin Expected Utility With Non-Unique Prior," *Journal of Mathematical Economics*, 18, 141–153.

Giraud, R. (2014): "Second Order Beliefs Models of Choice Under Imprecise Risk: Nonadditive Second Order Beliefs Versus Nonlinear Second Order Utility," *Theoretical Economics*, 9, 779–816.

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

Halevy, Y. and V. Feltkamp (2005): "A Bayesian Approach to Uncertainty Aversion," *The Review of Economic Studies*, 72, 449–466.

Halevy, Y., D. Walker-Jones, and L. Zrill (2023): "Difficult Decisions," *Working Paper.*

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.

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

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

Ke, S. and C. Zhao (2023): "From Local Utility to Neural Networks," *Working Paper.*

Klibanoff, P., M. Marinacci, and S. Mukerji (2005): "A Smooth Model of Decision Making Under Ambiguity," *Econometrica*, 73, 1849–1892.

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

Nau, R. F. (2006): "Uncertainty Aversion With Second-Order Utilities and Probabilities," *Management Science*, 52, 136–145.

Olszewski, W. (2007): "Preferences Over Sets of Lotteries," *The Review of Economic Studies*, 74, 567–595.

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

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

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

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

Savage, L. (1954): *The Foundations of Statistics*, New York: Wiley.

Schmeidler, D. (1989): "Subjective Probability and Expected Utility Without Additivity," *Econometrica: Journal of the Econometric Society*, 571–587.

Segal, U. (1987): "The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach," *International Economic Review*, 175–202.

——— (1990): "Two-Stage Lotteries Without the Reduction Axiom," *Econometrica: Journal of the Econometric Society*, 349–377.

Seo, K. (2009): "Ambiguity and Second-Order Belief," *Econometrica*, 77, 1575–1605.

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

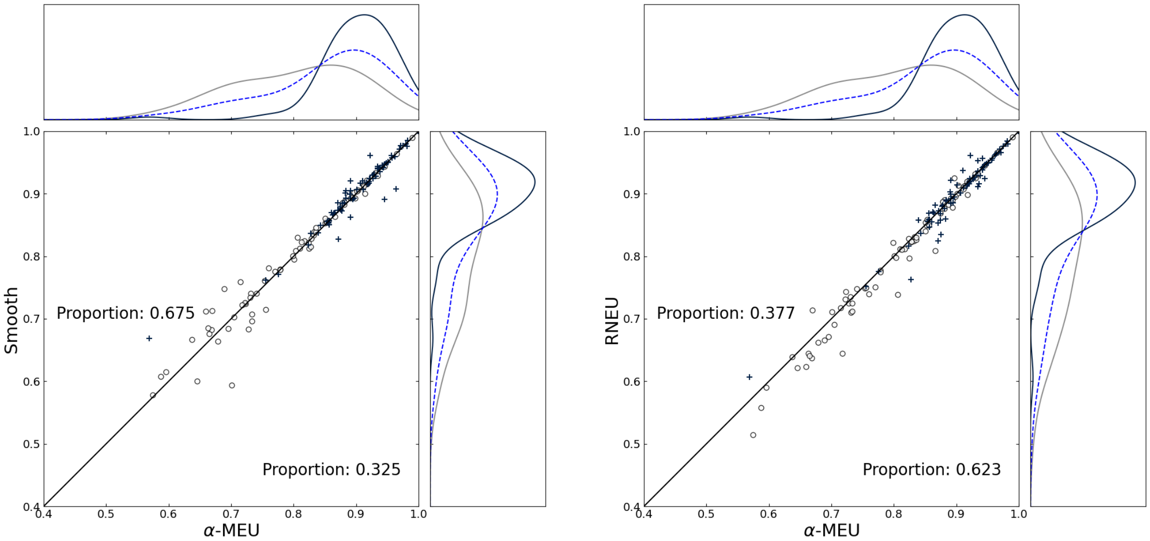
Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

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

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

Von Neumann, J. and O. Morgenstern (1947): "Theory of Games and Economic Behavior," in *Theory of Games and Economic Behavior*, Princeton University Press.

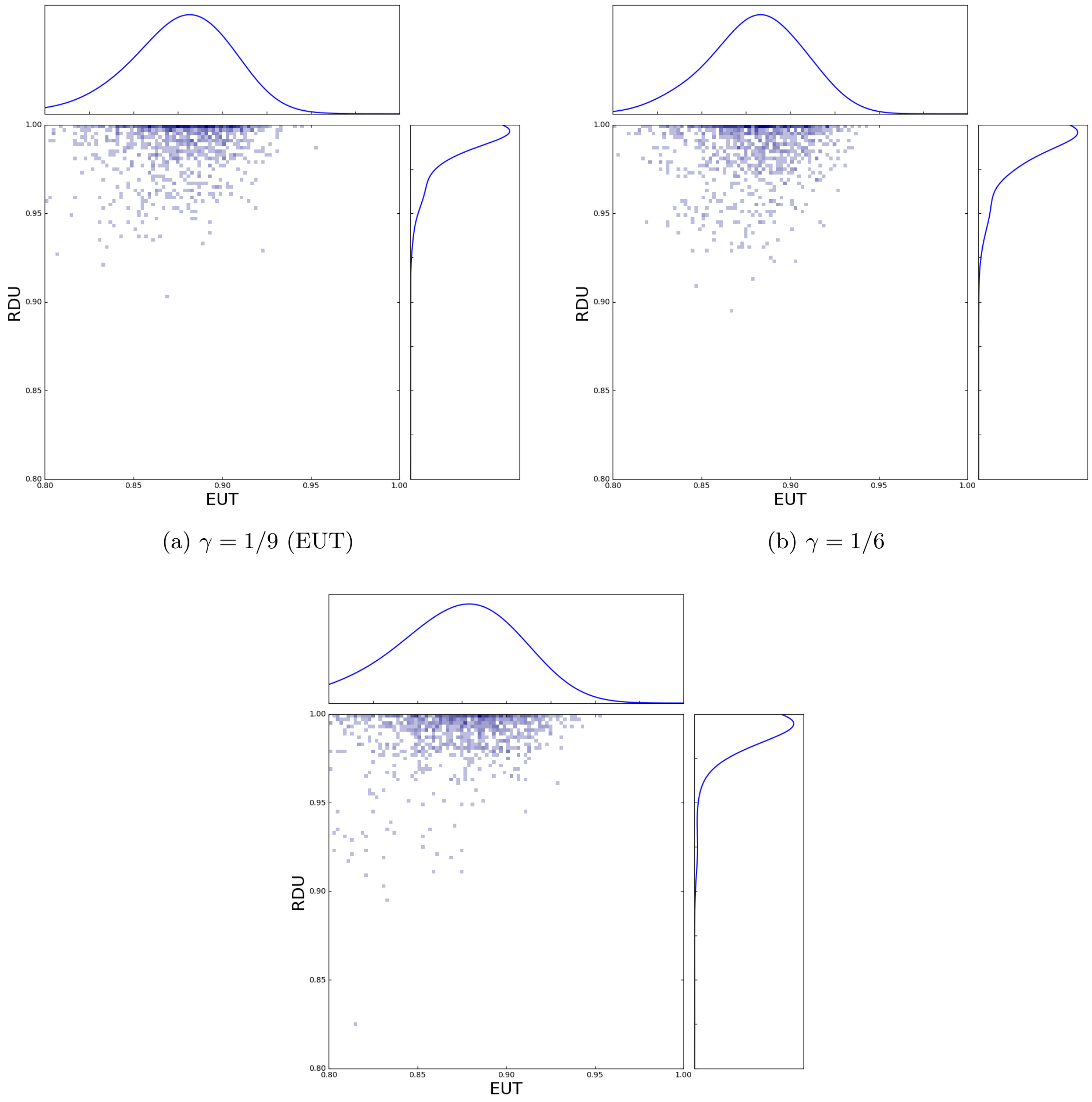
# A Additional Figures and Tables



(a) $\alpha$-MEU vs. Smooth Ambiguity

(b) $\alpha$-MEU vs. RNEU

Figure A.1: Scatterplot of completeness of standard economic models compared to nonstandard economic models. Subjects with below median CCEI are plotted as circles with a gray outline, and subjects with above median CCEI are plotted as navy "+" symbols. The light gray, navy, and dotted blue curves plot kernel density estimates of the marginal distributions for below median CCEI subjects, above median CCEI subjects, and all subjects, respectively.



(a) $\gamma = 1/9$ (EUT)

(b) $\gamma = 1/6$

(c) $\gamma = 1/3$

Figure A.2: Heatmap of completeness of EUT and RDU on simulated data with various levels of $\gamma$.

# B Analysis of "irrational" subject DGPs

First, consider the DGP that exclusively chooses portfolios that allocates all income to the most expensive Arrow security. Because both EUT and RDU satisfy monotonicity with respect to FOSD, there is a limit in "how close" the models can get to this "anti-risk neutral" behavior. Let $x_{(j)}$ denote the amount of investment in the $j$th cheapest Arrow security. It is clear that satisfying monotonicity with respect to FOSD requires that $x_{(1)} \leq x_{(2)} \leq x_{(3)}$. Dividing by the sum $x_1+x_2+x_3$ maintains the inequality. Hence, if a subject follows the "anti-risk neutral" decision rule, they set $\frac{x_{(1)}}{x_1+x_2+x_3} = 1$. The way to minimize the squared error subject to satisfying monotonicity, in terms of any two relative demands, is to set $x_{(1)} = x_{(2)} = x_{(3)} = 1/3$, which is achievable by both RDU and EUT models. Hence, the models will generate the same error for this DGP.

Next, consider the DGP that exclusively chooses to allocate all income to the $x$ account. For simplicity of analysis, assume that some price ratio $(p_1, p_2, p_3)$ is asked for all six possible permutations of states. Since EUT and RDU are both symmetric, let $\hat{d}_1$ and $\hat{d}_2$ denote the relative demands for the good with $p_1$ and $p_2$, respectively. For each of the six permutations, the errors are shown in Table B.2.

| Permutation | Relative Demand | Error |
|---|---|---|
| 123 | $(\hat{d}_1, \hat{d}_2)$ | $(1-\hat{d}_1)^2 + (0-\hat{d}_2)^2$ |
| 132 | $(\hat{d}_1, 1-\hat{d}_1-\hat{d}_2)$ | $(1-\hat{d}_1)^2 + (\hat{d}_1+\hat{d}_2-1)^2$ |
| 213 | $(\hat{d}_2, \hat{d}_1)$ | $(1-\hat{d}_2)^2 + (0-\hat{d}_1)^2$ |
| 231 | $(\hat{d}_2, 1-\hat{d}_1-\hat{d}_2)$ | $(1-\hat{d}_2)^2 + (\hat{d}_1+\hat{d}_2-1)^2$ |
| 312 | $(1-\hat{d}_1-\hat{d}_2, \hat{d}_1)$ | $(\hat{d}_1+\hat{d}_2)^2 + (0-\hat{d}_1)^2$ |
| 321 | $(1-\hat{d}_1-\hat{d}_2, \hat{d}_2)$ | $(\hat{d}_1+\hat{d}_2)^2 + (0-\hat{d}_2)^2$ |

Table B.2: Estimation error for each permutation of prices $(p_1, p_2, p_3)$.

The mean squared error is thus $\frac{1}{3}\left(4\hat{d}_1^{\,2} + 4\hat{d}_1\hat{d}_2 - 4\hat{d}_1 + 4\hat{d}_2^{\,2} - 4\hat{d}_2 + 3\right)$, which is minimized when $\hat{d}_1 = \hat{d}_2 = \frac{1}{3}$. This again is achievable by both RDU and EUT.