Scaling Up:
Individual-Level Transfer Performance of Models

Keaton Ellis, Shachar Kariv, and Erkut Y. Ozbay*
Latest draft here

November 1, 2024

Abstract

This study investigates the transferability of economic models for individual decision-making across different risk domains, specifically comparing performance between two- and three-state budgetary environments. Utilizing within-subject laboratory data, we evaluate the ability of Expected Utility Theory (EUT), Disappointment Aversion (DA), and machine learning models to predict choices when estimated in a simpler two-state environment and applied to a more complex three-state environment at the individual level. Our findings reveal two key insights: (i) there is substantial transferability across domains for the vast majority of subjects; and (2) EUT demonstrates substantial transferability, maintaining approximately 92.9% of its within-domain predictive accuracy when generalized across domains, outperforming both DA and machine learning models in terms of predictive consistency. These results underscore the robustness of parsimonious economic models, particularly EUT, in providing reliable extrapolations across experimental contexts, suggesting their utility in applications where predictions span diverse risk settings.

*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

Economic theory seeks to identify the core concepts underlying choice behavior, which can then be modeled and applied to a multitude of economically-relevant scenarios. However, oftentimes the data that is collected to create an economic model is meaningfully different from the scenarios which an analyst wants to generate predictions. These scenarios require models to be not just accurate out-of-sample prediction within a domain, but across domains. For example, financial analysts may provide prospective investors with mock investment scenarios to elicit risk preferences, which then guide retirement planning scenarios. It is natural to quantify the extent to which model predictions transfer across relevant, but separate, domains.

More concretely, consider a choice under risk experiment with linear budget sets. There are a set of states of nature S, with each state being equally likely. Subjects select portfolio allocations of Arrow securities corresponding to the states, where an Arrow security is defined to be a promise to deliver one dollar if state $s \in S$ occurs and nothing otherwise. Let the vector $\mathbf{x}$ denote a portfolio of securities, where each component 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 $\mathbf{p} \geq 0$.

Furthermore, consider two sizes of S: $|S| = 2$ and $|S| = 3$. While these environments appear similar, there is a large jump in richness between them, importantly with regards to generalizations of expected utility theory (EUT). Theoretical properties, such as singleton separability and joint separability, may also be equivalent in the former case but not in the latter. Thus, the three-state experiment adds much of the richness and complexity also present for larger values of $|S|$.

We use these two environments to empirically determine the transfer performance of economic models from the case of $|S| = 2$ to $|S| = 3$. In particular, we evaluate the out-of-domain performance of expected utility (EUT) against three metrics. First, we benchmark out-of-domain performance with "within-domain, out-of-sample" performance. This exercise identifies the extent to which behavior shifts between domains, and usefully acts as an upper bound in the prediction quality of out-of-domain performance (Andrews et al., 2023). Second, we replicate the same out-of-domain exercise

with disappointment aversion (DA, Gul, 1991). These models extend expected utility theory by relaxing the independence axiom to allow for behaviors such as nonlinear interpretation of probabilities and the Allais (1953) paradox, respectively. This exercise evaluates whether the additional parameter of disappoint aversion stores more meaningful information about a subject's choices over the single parameter we assume for expected utility, or whether they store meaningful information at all.

Finally, we conduct an out-of-domain exercise with black box machine learning models. These large, flexible model classes are designed for function approximation via large quantities of data. In the

Prior work has benchmarked economic models against machine learning models in out-of-sample prediction tasks, with mixed results depending on level of analysis (individual-level vs. pooled) and data environment (choice under risk, initial play in matrix games, etc.). However, the message is straightforward: if the machine learning model outperforms the economic model, then there is a regularity in choice behavior that the machine learning model captures that economic models do not capture; it is then the analysts' job to understand that regularity and incorporate it into the theory (Fudenberg and Liang, 2019).

The out-of-domain, on the other hand, is subtly more difficult. An economic model of choice is easily portable between these environments and many others because we define the economic model over a larger space than a single experiment. However, machine learning models are inherently restricted to a single input space dependent on training data, which changes when moving from two dimensions to three. Hence, the machine learning model needs additional information about mapping input spaces, which we make in the form of a symmetry assumption. Without such impositions, there is no reason to expect high quality extrapolation. For example, Andrews et al. (2023) find that machine learning models fail to predict certainty equivalences that are slightly outside the domain of training data.

To evaluate model performance, we use a measure of *transfer completeness* adapted from Fudenberg et al. (2022), which scales the out-of-sample predictive power of a model to that of a naive benchmark and a model with "irreducible" error. We analyze transfer performance of models using two estimates of irreducible error: perfect

prediction and within-domain, out-of-sample prediction, as described above.

Our analysis yields two main results:

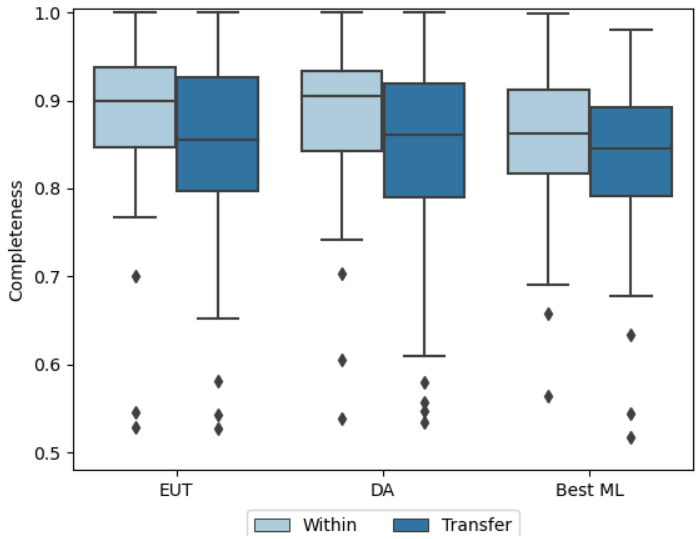


Figure 1: The average within completeness and transfer completeness of models

Mild transfer loss: First, data from two-state allocations is an excellent proxy for choice behavior in three-state allocations, and we observe a tight relationship between transfer performance and within sample performance. Figure 1 shows a box plot of within-domain and transfer completeness of EUT, disappointment aversion, and the subjectwise most complete machine learning model. The line in the box represents the median, the box represents the middle 50% of the distribution, and the whiskers represent the ends of the distribution sans outliers, which for each model are plotted separately below.

As expected, the transfer performance of all models is lower than that of within-domain estimation. However, the degree to which it lowers is marginal: for EUT, the average transfer completeness relative to within-domain completeness is 92.9%, with similar or better transfer results for the other models considered.[1] Hence, much

[1]The same quantities for disappointment aversion and Best ML are 92.5% and 98.5%, respec-

of a model's predictable variation in a three-good environment is transferable from models estimated on a two-good environment, despite the added complexity.

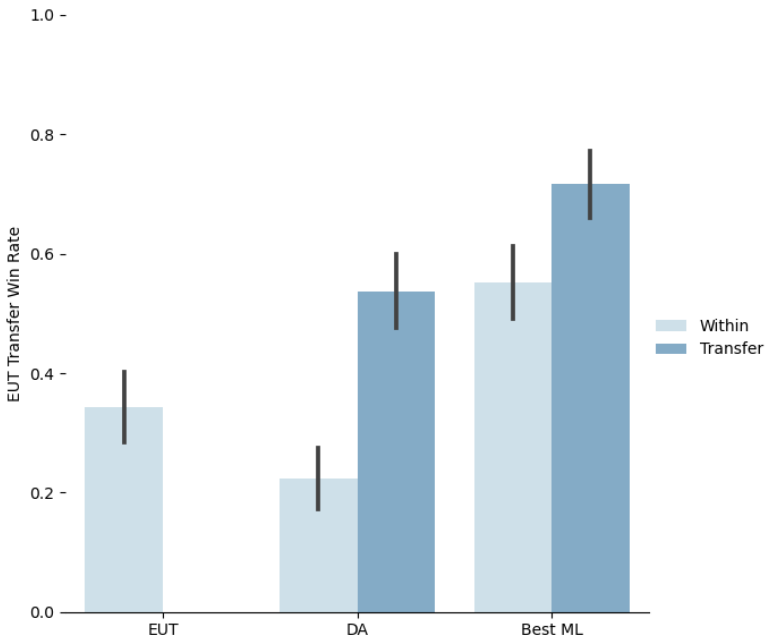


Figure 2: The proportion of subjects for whom transfer-EUT is more complete, by model and estimation method.

Strong EUT transferability: Second, the single parameter of EUT is sufficient for transferring demand information from two-state allocations to three-state allocations. Figure 2 shows the proportion of subjects for whom transfer-EUT is more complete, which we denote the "win rate" of EUT transfer, for each model and for each estimation method. We see two key results.

Unsurprisingly, for EUT, within-domain estimation is significantly more likely to be more complete than transfer estimation (binomial test, $p = 0.01$). The same is true for within-domain DA. However, the win rate of transfer estimation of EUT is approximately even with within-domain Best ML ($p = 0.396$), despite being trained on the three-good experimental data. When comparing transfer performances, the

tively.

win rate of EUT is approximately that of DA ($p = 0.545$), but is significantly higher than 0.5 for Best ML ($p < 0.001$). Hence, the additional flexibility of these models does not systematically afford increased transferability across domains relative to EUT.

We additionally select subjects to highlight the extent to which economic models can transfer. We find the subjects with the highest deviance comparing within-domain and between-domain completeness stem from vastly different choice patterns between two-state allocations and three-state allocations.[2] Additionally, we find that economic models can assign an incorrect heuristic to choices in two-states, which may lead to arbitrarily large errors in three states. Finally, we examine the subjects for whom EUT performs worst compared to machine learning models. We find that the increased performance seems coincidental: it stems from subjects exhibiting risk neutral behavior in two states and smooth downward sloping demand in three states. These subjects are best predicted by linear regularized regressions, whose imperfect fit of risk neutral behavior in two states coincidentally aligns with the behavior in three states. Other machine learning models such as random forests, gradient-boosted regression trees, exhibit equally poor performance as EUT. Thus, we believe the results stem less from a regularity in choice data and more from a coincidence in model behavior.

2 Related Literature

Our paper closely relates to papers that evaluate the transfer performance of economic models in various settings. Andrews et al. (2023) conduct an aggregate-level exercise to predict certainty equivalences of binary lotteries and evaluate how well the model transfers to different distributions of problems and accompanying data sets of subject pools that face those distributions. They find that economic mod-

[2] At least, there are no decision rules we are aware of that simultaneously generate these types of choice patterns in two-state and three-state experiments, although frequently the choice pattern in the experiments can be easily explained standalone.

els transfer approximately equally well across subject pools, and do so better than machine learning models. However, they find that when controlling for the distribution of problems, the machine learning models generate similar transfer performance estimates as economic models.

Kobayashi and Lucia (2023) compare the relative performance of EUT, CPT, and machine learning models in binary prediction tasks in two scenarios with prevalent non-EUT behavior: common ratio tasks á la Allais (1953) and preference randomization tasks á la Agranov and Ortoleva (2017). They find that machine learning models have better out-of-sample prediction than three-group economic mixture models within a task type, but worse cross-task prediction when training on questions of one type and testing on another. Fehr et al. (2023) uses a Bayesian nonparametric clustering method to identify clusters of preference types in social preference space. For estimation, they conduct a similar exercise to Kobayashi and Lucia (2023) by estimating models on 12 "centered" budget lines that contain an equal allocation and making predictions on up to 52 "non-center budget lines" which do not contain an equal allocation (or contains it at an endpoint of a budget line). They show that the model produces similar estimates to individual-level preference estimation and outperforms machine learning models, justifying their method over individual-level estimation due to parsimony. Relative to these two papers, our task focuses instead of increasing the complexity of the environment subjects face, rather than modifying the payoff values within a given environment. Additionally, we provide machine learning models with some information on how to transfer, and find smaller gaps in performance between economic and machine learning models.

We employ similar methods. Our discussion of transfer performance relative to within performance can be interpreted in the Andrews et al. (2023) framework as "transfer deterioration", which is also similar to the completeness of transfer estimation relative to random uniform choice and within-sample estimation (Fudenberg et al., 2022). However, our environment differs in two important ways. First, our transfer question of interest focuses not on a change in distributions of choices conditional on problems, but instead on a change in the covariates between domains. In this sense, we are interested in how the same decision maker makes decisions in

new scenarios, as opposed to how different decision makers approach the same environment. Prior papers sidestep the issue of various inputs by evaluating differences in distribution (holding covariates fixed) or explicitly use this limitation to showcase changes in distribution of choices. Second, our rich data sets allow us to performance within-subject analysis at the individual level, as opposed to aggregate-level between-sample. By doing so, we produce a strong evaluation of transfer performance of economic models.

We also relate to papers that compare the out-of-sample prediction performance of economic models against machine learning models, but do not vary the choice environment. These papers span much of economic theory, including choice under risk and ambiguity (Peysakhovich and Naecker, 2017, Plonsky et al., 2019, Peterson et al., 2021, Fudenberg and Puri, 2022, Hsieh et al., 2023, Ke and Zhao, 2023, Ellis et al., 2023), preference elicitation methods (Clithero et al., 2023), and initial play in matrix games (Fudenberg and Liang, 2019), among the topics discussed above. Also of note are papers such as Halevy et al. (2018), which employ prediction to evaluate various structural estimation methods for models of choice under risk.

Our paper also touches upon the consistency of experimental risk elicitation methods. Friedman et al. (2022) generate risk preference estimates for six elicitation methods, evaluating the correlation between the estimates based on design features. They find that designs align stronger when attributes of the design, such as the number of alternatives or whether the representation of choices is visual, align. Much of the remaining literature analyzing elicitation methods hold an elicitation method fixed and analyze changes in behavior and/or incentive compatibility when varying attributes of the method.[3] While our analysis is instead focused on prediction and not on incentive compatibility or recover of preferences, we are similarly interested in the portability of economic models across different experimental environments.

[3]For examples, Brown and Healy (2018) evaluates multiple price list behavior when showing questions on one screen versus each question appearing on a separate screen. Cox et al. (2015) focus on varying the payment mechanism for a sequence of decision problems.

3 Template for Analysis

3.1 Experimental design

Our dataset is comprised of data from laboratory experiments in which 67 subjects solve a portfolio choice problem. The data was collected previously; 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, see Choi et al. (2007) and Dembo et al. (2021).

Subjects participated in two experiments. In the first experiment, denoted the "2D experiment", there are two equiprobable states of nature $s = 1, 2$ 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 $\mathbf{x} = (x_1, x_2) \geq \mathbf{0}$ 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 $\mathbf{p} = (p_1, p_2) \geq \mathbf{0}$ is the vector of security prices and p_s denotes the price of security s. In the second experiment, denoted the "3D experiment", there are instead three equiprobable states $s = 1, 2, 3$, portfolios consist of three accounts $\mathbf{x} = (x_1, x_2, x_3) \geq \mathbf{0}$ and must satisfy a budget constraint with three prices $\mathbf{p} = (p_1, p_2, p_3) \geq \mathbf{0}$.

3.2 Evaluating model performance

We conduct two exercises. First, we evaluate models using a within-domain, out-of-sample prediction task for the 3D environment. We use the completeness measure from Fudenberg et al. (2022), which reports the fraction of predictable variation a model captures. Second, we evaluate models with a between-domain prediction task, where we estimate a model on data from the 2D environment and then predict choices in the 3D environment. We again use completeness to report the fraction of predictable variation a model captures between domains. To distinguish, we refer to the former as "Within completeness" and the latter as "Transfer completeness". Our main analysis compare within vs. transfer completeness for a given model and

compare transfer completeness across models.

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 $\mathbf{x}$, let 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 one state is sufficient to know demand in the 2D environment, and knowing the relative demand for two states is sufficient to know demand in the 3D environment. We use $\mathbf{d} = d_1$ and $\mathbf{d} = (d_1, d_2)$ for the 2D and 3D environments, respectively. Thus, in all analysis below, the error 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 squared error *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 assigned to a predicted relative demand $f(\mathcal{B}^i)$ when the actual relative demand is $\mathbf{d}^i$.[4] 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 $(\mathcal{B}, \mathbf{d})$.[5] We are interested in comparing set of mappings parametrized by some Θ, $\mathcal{F}_\Theta = \{f_\theta\}_{\theta \in \Theta}$, 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:

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

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

[5] 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.

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

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 κ_Θ, is defined by

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

where f_n is a naive benchmark mapping and the 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 $\mathbf{d}$. Given demand $\mathbf{d}$, the expected error is $\frac{1}{3}(1-3\mathbf{d}+3\mathbf{d}^2)$ in the 2D environment and $\frac{1}{3}-\frac{2}{3}d_1-\frac{2}{3}d_2+d_1^2+d_2^2$ in the 3D environment.[6]

[6]In the 2D environment, this is analogous to a uniform choice over the interval $[0,1]$. Given a subject's actual choice $\mathbf{d}$, the error of a naive model is $\ell(\mathbf{d}_n, \mathbf{d}) = (\mathbf{d}_n - \mathbf{d})^2$. If $\mathbf{d}_n \sim U[0,1]$, then the expected MSE conditional on the actual choice $\mathbf{d}$ is:

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

In the 3D environment, 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 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:

Within-domain We estimate $\mathcal{E}_P(f_n)$ and $\mathcal{E}_P(f^*_\Theta)$ using data from the 3D environment for each subject, and then plug these estimates into the formula for κ_Θ 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^*_Θ as $\hat{f}^{-k} = \arg\min_{f_\theta \in \mathcal{F}_\Theta} \frac{1}{45}\sum_{i \notin Z_k} \ell(f_\theta(\mathcal{B}^i), \mathbf{d}^i)$. 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 κ_Θ is normal.

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

Between-domain We estimate $\mathcal{E}_P(f^*_\Theta)$ for the 3D environment by minimizing loss on 2D data. Let $\{(\mathcal{B}^i, \mathbf{d}^i)\}_{i=1}^{50}$ denote data in the 3D environment, $\{(\mathcal{B}^j, \mathbf{d}^j)\}_{j=1}^{50}$ denote data in the 2D environment. f^*_Θ is estimated using

$$\hat{f}^*_\Theta = \arg\min_{f_\theta \in \mathcal{F}_\Theta} \frac{1}{50} \sum_{j=1}^{50} \ell\left(f_\theta(\mathcal{B}^j), \mathbf{d}^j\right)$$

The estimate of model error $\mathcal{E}_P(f^*_\Theta)$, denoted $\hat{\mathcal{E}}_\Theta$, is the loss of $\hat{f}^*_\Theta$ in the 3D environment:

$$\hat{\mathcal{E}}_\Theta = \frac{1}{50} \sum_{i=1}^{50} \ell\left(\hat{f}^*_\Theta(\mathcal{B}^i), \mathbf{d}^i\right)$$

The estimate of $\mathcal{E}_P(f_n)$ is identical to that of the within-domain estimation.

3.3 Economic models

We consider expected utility theory (EUT) and disappointment aversion (DA) by Gul (1991). Since EUT is a special case of DA, we describe DA in detail below and note how it nests EUT.

Disappointment Aversion In the disappointment aversion model, a (simple) lottery $\mathcal{L}$ is evaluated by placing differing weights on "elating" outcomes, which are preferred to the lottery, and "disappointing" outcomes, which are not preferred to the lottery. Thus, a lottery is represented as an elation/disappointment decomposition (EDD), denoted (α, q, r), which creates two lotteries q and r containing the more preferred and less preferred alternatives of a lottery, respectively, such that $\alpha q + (1-\alpha)r = \mathcal{L}$. Given an EDD (α, q, r), the utility function of disappointment aversion is

$$U(\tilde{\mathbf{x}}) = \gamma(\alpha) \sum_x u(x)q(x) + [1 - \gamma(\alpha)] \sum_x u(x)r(x),$$

where $\gamma(\alpha) = \frac{\alpha}{1+(1-\alpha)*\beta}$, $\beta \geq -1$, is the weight placed on elating outcomes, and $\tilde{X}$ is a rank-ordered profile. For $\beta > 0$, $\gamma(\alpha) \leq \alpha$ and thus additional weight is placed

on disappointing outcomes. The opposite is true for $\beta < 0$. EUT is nested as the special case of $\beta = 0$, which shuts down the $\gamma(\alpha)$ distortion channel.

In the 2D environment with two equiprobable states, the model condenses to:

$$U(\tilde{\mathbf{x}}) = \gamma(0.5)u(x_H) + [1 - \gamma(0.5)]u(x_L),$$

where again $\tilde{\mathbf{x}} = (x_L, x_H)$ is a rank-ordered portfolio with $x_L < x_H$. The estimated β value is obtained by solving $\gamma(0.5) = \frac{1/2}{1+1/2*\beta}$, which results in $\beta = \frac{1}{\gamma(0.5)} - 2$.

In 3D, disappointment aversion predicts two separate utility values depending on whether the EDD is $(\frac{1}{3}, x, \frac{1}{2}y + \frac{1}{2}z)$ or $(\frac{2}{3}, \frac{1}{2}x + \frac{1}{2}y, z)$. The utility of a rank-ordered portfolio (x_L, x_M, x_H) is

$$U(\tilde{\mathbf{x}}) = \begin{cases} \frac{1}{3+2\beta}\left[u(x_H) + (1+\beta)u(x_M) + (1+\beta)u(x_L)\right] & EDD = (\frac{1}{3}, x_H, \frac{1}{2}x_M + \frac{1}{2}x_L) \\ \frac{1}{3+\beta}\left[u(x_H) + u(x_M) + (1+\beta)u(x_L)\right] & EDD = (\frac{2}{3}, \frac{1}{2}x_H + \frac{1}{2}x_M, x_L) \end{cases}$$

The first case occurs when $U(\tilde{\mathbf{x}}) > u(x_M)$, and the second case occurs when $U(\tilde{\mathbf{x}}) < u(x_M)$. The boundary condition is thus $U(\tilde{\mathbf{x}}) = u(x_M)$, which occurs when $u(x_H) + (1+\beta)u(x_L) = (2+\beta)u(x_M)$.

Each sub-function of the piecewise function $U(\tilde{\mathbf{x}})$ is a special case of RDU, but disappointment aversion is not a special case of RDU. Within the RDU model, setting $w_1 = \frac{1+\beta}{3+2\beta}$ and $w_2 = \frac{2+2\beta}{3+2\beta}$ results in disappointment aversion utility for $EDD = (\frac{1}{3}, x, \frac{1}{2}y + \frac{1}{2}z)$. Setting $w_1 = \frac{1+\beta}{3+\beta}$ and $w_2 = \frac{2+\beta}{3+\beta}$ results in disappointment aversion utility for $EDD = (\frac{2}{3}, \frac{1}{2}x + \frac{1}{2}y, z)$.

Implementation For each subject, for within-domain and between-domain estimation, and for disappointment aversion and EUT, estimation is done via nonlinear least squares. We estimate using two Bernoulli utility functions $u(x)$. First, we use the constant absolute risk aversion (CARA) utility function $u(x) = -e^{-Ax}$, with

$A \geq 0$. Second, we use the constant relative risk aversion (CRRA) function

$$u(x) = \begin{cases} \frac{x^{1-\rho}}{1-\rho} & \rho \neq 1 \\ \log(x) & \rho = 1 \end{cases},$$

with no limitation on ρ. When reporting the performance of a model, we report the more complete of CARA or CRRA.

3.4 Machine learning models

We utilize a suite of machine learning models to compare against EUT and DA across three classes regularized regressions, tree-based, and neural networks. We briefly cover each type of model below, referring readers to Hastie et al. (2009) for more in-depth discussion.

Regularized regressions select from the class of linear models $f_{Lasso}(\mathcal{B}) = \hat{\beta}^T \mathcal{B}$ in the same manner as OLS, but do so with a modified objective function that adds the norm of the coefficients $\lambda \|\beta\|_p$ as a "regularization" term, multiplied by a tuning parameter λ controlling the strength of the regularization. We use $p = 1$ and $p = 2$, which correspond to lasso and ridge regression, respectively. We use leave-one-out cross-validation to determine the parameter $\lambda \in [0, 0.2, 0.4, 0.6, 0.8, 1]$.

Tree-based models select from the class of piecewise step functions. Starting from the entire input space, a single decision tree will recursively partition the input space in a binary, locally optimal fashion to minimize prediction error of separate values predicted for each subset. Since this partitioning process, if allowed to continue without restraint, would end with each data point in its own partition with perfect within-sample prediction, we intentionally limit the depth of trees and set a minimum number of observations per terminal partition.

Since a single decision tree is typically insufficient to express rich variation, we utilize two tree aggregation methods as well. Random forests (Breiman, 2001) average across multiple trees trained on bootstrapped samples with a subset of variables. Gradient boosted decision trees iteratively train trees on weighted residuals of the

previous trees, whose final prediction rule aggregates previous trees. In particular, this model is commonly seen as best model for tabular data (Schuler et al., 2023).

Finally, we utilize neural networks, which iteratively conduct affine transformations and nonlinear transformations on the input space, eventually aggregated as a demand prediction. We implement multilayer perceptron neural networks using specifications from Hsieh et al. (2023). We set the nonlinear transformation to be $\sigma(x) = \max\{x, 0\}$ and 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.

Note that, regardless of the model used, there is an inherent nontransferability of the machine learning models when estimating on the 2D environment and predicting on the 3D environment because the training data is of a different input size than the testing data. To match the two, we input the 2D data as coming from a 3D environment where one price is infinitely large. Hence, a 2D budget set $\mathcal{B}$ is input as the vector $[p_1, p_2, 0]$, the chosen portfolio is $\mathbf{x} = [x_1, x_2, 0]$ and the demand is $\mathbf{d} = [\frac{x_1}{x_1+x_2}, \frac{x_2}{x_1+x_2}]$. This is known as "zero-padding", and is a simple and common method to maintain input size.

Without any further modifications, training a machine learning model on this input would perform poorly, as there is no variation in the third component. To ensure that a machine learning model learns relationships between all three states, we impose a symmetry assumption. We create six copies of the data and label the copies according to the six possible permutations of accounts x, y, and z. Thus, a single portfolio choice that a subject made from a budget set $\mathcal{B}$ is input six times as:

$\mathbf{p}$	$\mathbf{x}$	$\mathbf{d}$
$[1/p_1, 1/p_2, 0]$	$[x_1, x_2, 0]$	$[\frac{x_1}{x_1+x_2}, \frac{x_2}{x_1+x_2}]$
$[1/p_1, 0, 1/p_2]$	$[x_1, 0, x_2]$	$[\frac{x_1}{x_1+x_2}, 0]$
$[1/p_2, 1/p_1, 0]$	$[x_2, x_1, 0]$	$[\frac{x_2}{x_1+x_2}, \frac{x_1}{x_1+x_2}]$
$[1/p_2, 0, 1/p_1]$	$[x_2, 0, x_1]$	$[\frac{x_2}{x_1+x_2}, 0]$
$[0, 1/p_1, 1/p_2]$	$[0, x_1, x_2]$	$[0, \frac{x_1}{x_1+x_2}]$
$[0, 1/p_2, 1/p_1]$	$[0, x_2, x_1]$	$[0, \frac{x_2}{x_1+x_2}]$

4 Results

Recall that Section 1 summarizes our main results: (i) there is mild transfer loss between two states and three states (see Figure 1), and (ii) EUT transfers at least as well as disappointment aversion and better than machine learning models (see Figure 2). We further investigate these claims in this section.

Table 1 provides a population-level summary of our results, elaborating on the information provided in Figures 1 and 2 above. The left column reports the average completeness of each model for each estimation method. The next block of five columns provides a five-number summary of average completeness, which are reported in Figure 1. The seventh column shows the win rate of EUT transfer against models, which are reported exactly in Figure 2. The final two columns show the average and median ratio of transfer completeness to within-sample completeness. The table elaborates on the first claim regarding distribution tightness. While the distributions of EUT within and EUT transfer are mildly significantly different (Kolmogorov-Smirnov test, $p = 0.043$), they are not for disappointment aversion (KS-test, $p = 0.329$) or machine learning (KS-test, $p = 0.447$). For EUT, the significance appears to stem from a stronger left skew in transfer completeness than within, since the majority of subjects possess a transfer completeness of at least 98.5% relative to within.

Next, Figure 4 shows scatter plots transfer completeness and within-domain completeness for EUT. The transfer completeness is plotted on the x-axis, and the within-domain completeness is plotted on the y-axis. The proportion of subjects with higher transfer completeness is shown below the diagonal, and the proportion of subjects with higher within-domain completeness is shown above the diagonal. Finally, we provide a kernel density estimate showing the marginal distribution of completeness scores for each estimation. The results, perhaps unsurprisingly, exhibit asymmetry in favor of within-domain. While 34.3% of subjects have higher transfer completeness than within-domain completeness, these subjects lie within a tight bound around the 45-degree line. On the contrary, the difference in completeness for the remaining 65.7% of subjects with a higher within-domain completeness than transfer complete-

Table 1: The transfer and within completeness of models

Model		$\hat{\kappa}$ Mean	Min	25%	Median	75%	Max	$Win\%_{EUT}$	$\hat{\kappa}_{transfer}/\hat{\kappa}_{within}$ Mean	Median
EUT	Within	88.3%	52.9%	84.7%	89.9%	93.8%	100.0%	34.3%	92.9%	98.5%
	Transfer	82.1%	-27.3%	79.6%	85.5%	92.7%	100.0%	-		
DA	Within	88.5%	53.9%	84.3%	90.5%	93.4%	100.0%	22.4%	90.9%	98.0%
	Transfer	80.6%	-28.2%	79.0%	86.1%	92.0%	100.0%	53.7%		
ML	Within	85.1%	45.4%	81.6%	86.3%	91.3%	99.9%	55.2%	98.5%	99.6%
	Transfer	83.4%	51.7%	79.1%	84.5%	89.2%	98.0%	71.6%		

The left column reports the average completeness of each model for each estimation method. The next block of five columns provides a five-number summary of average completeness, which are reported in Figure 1. The seventh column shows the win rate of EUT transfer against models, which are reported exactly in Figure 2. The final two columns show the average and median ratio of transfer completeness to within-sample completeness.

ness is significantly more heterogeneous. These results are also robust to analyzing non-EUT models of choice under risk. Appendix Figure A.1 plots transfer completeness against within-domain completeness for disappointment aversion and machine learning. Both panels are similar to Figure 4, but exhibit higher levels of dispersion.

Next, we examine the transfer completeness between models. Figure 5 plots the transfer completeness of EUT against disappointment aversion and machine learning in the same manner as in Figure 4. Panel 5a plots EUT and disappointment aversion, and Panel 5b plots EUT and machine learning.

First, disappointment aversion and EUT exhibit approximately equal completeness. EUT is more complete than disappointment aversion for only 53.7% of subjects, which is not significantly different than 50% (binomial test, $p = 0.313$). Additionally, there are very few subjects with completeness scores away from the 45-degree line, indicating a consistency between model predictions. Second, EUT on average outperforms machine learning models, and is more complete for 71.6% of subjects (binomial test, $p < 0.01$). Compared to the relationship between EUT and disappointment aversion, there is more dispersion primarily in favor of EUT, but also

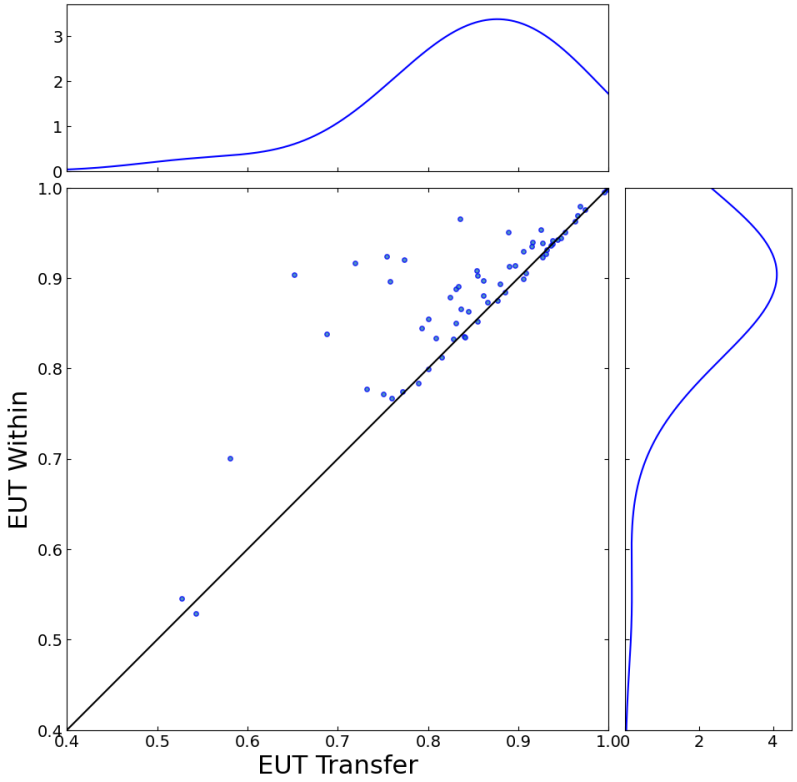
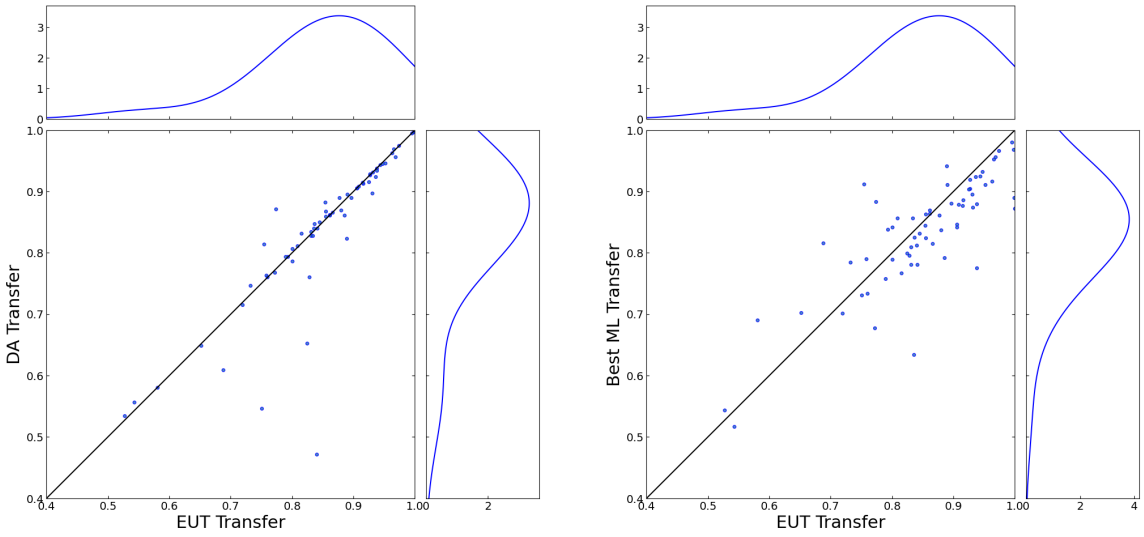


Figure 3: Completeness

Figure 4: EUT Transfer completeness relative to EUT within-domain completeness



(a) EUT vs. disappointment aversion

(b) EUT vs. ML

Figure 5: Relative transfer completeness of EUT

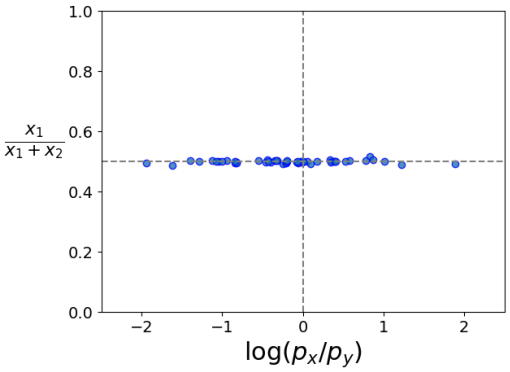
some in favor of machine learning models. Not displayed are subjects exhibiting extreme low levels of EUT transfer completeness, but reasonable machine learning completeness. We examine such a subject in the next section.

4.1 Individual subject analysis

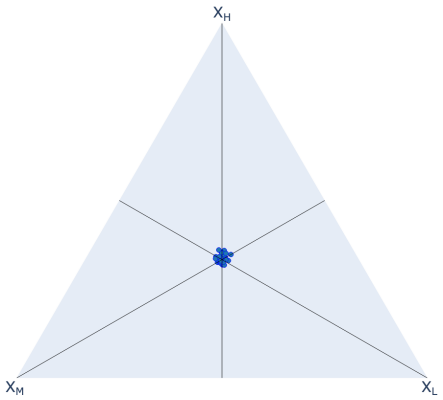
Finally, we examine the choices of select subjects that demonstrate the extent to which economic models can transfer. Here, we highlight two general classes - (i) strong transferability within the scope of utility maximization and (ii) strong transferability outside the scope of utility maximization.

Utility-maximizers Figure 6 shows the 2D and 3D data for subjects 101, 221, and 222. The left-hand column shows the relative demand in 2D for the three subjects. The x-axis corresponds to the log price ratio and the y-axis corresponds to the relative demand $\mathbf{d}$ for account x. The right-hand column shows choices plotted in a rank-ordered budget set, with the most expensive account in the bottom right, the least expensive account in the top, and the final account in the bottom left.

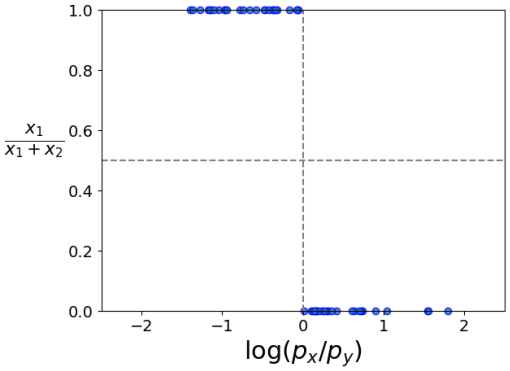
Figure 6: Demand of subjects 101, 221, and 222 for 2D and 3D.



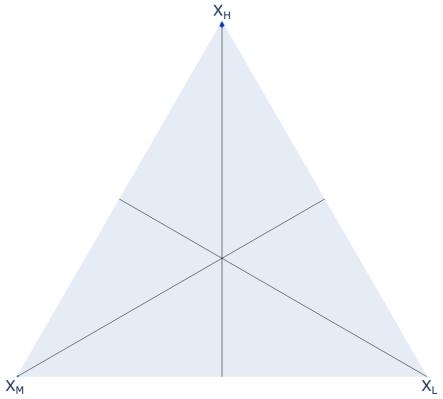
(a) ID 101 2D



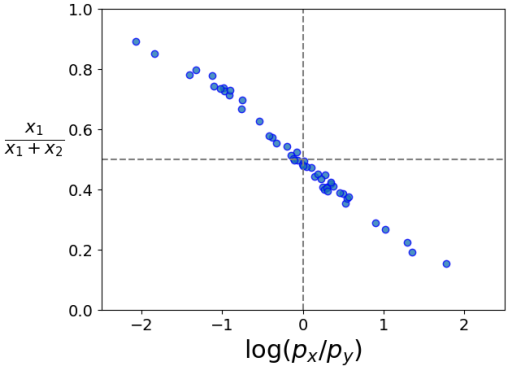
(b) ID 101 3D



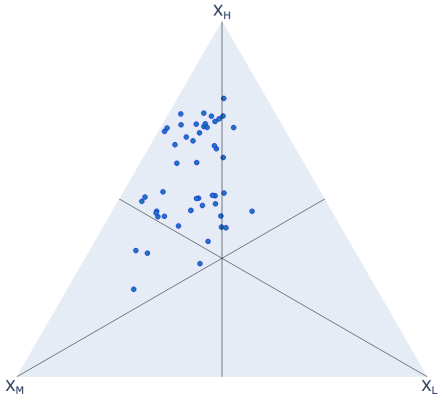
(c) ID 221 2D



(d) ID 221 3D



(e) ID 222 2D



(f) ID 222 3D

The left-hand column shows the relative demand in 2D for the three subjects. The x-axis corresponds to the log price ratio and the y-axis corresponds to the relative demand **d** for account x. The right-hand column shows choices plotted in a rank-ordered budget set, with the most expensive account in the bottom right, the least expensive account in the top, and the final account in the bottom left.

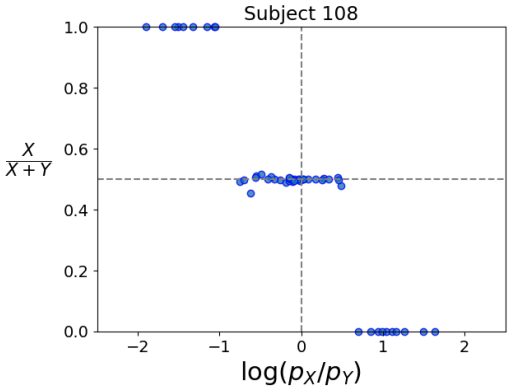
Subject ID 222 consistently chooses the middle of the 45 degree line in 2D, which is consistent with log utility preferences. Their results in the 3D environment exhibit a similar smoothness, albeit at a slightly less risk averse rate - for comparison, the estimate of ρ for CRRA is 1.001 in the 2D environment and 0.645 in the 3D environment. Subject ID's 101 and 221, who exhibit heuristics of choosing the intersection with the 45-degree line and the highest intercept, respectively. These choice rules are consistent with (infinitely) risk averse and risk neutral preferences, respectively. Both subjects exhibit a high completeness in all settings; for subject ID 221, transfer and within-domain EUT completeness are 96.3% and 96.2%, respectively. For subject ID 101, EUT transfer and within-domain completeness are 99.96% and 99.9999% respectively.

Consistent, but not utility maximizing Figure 7 plots plots the data for subjects 108 and 207. Subject 207 exclusively invests in the asset with higher intercept, which is consistent with risk neutral preferences. Subject 108 exhibits the same behavior for extreme price ratios, but instead invests equally in the two assets when prices are relatively close. This is consistent with disappointment averse, risk neutral preferences. Despite this good fit in two states, the disappointment aversion transfer completeness is 47.1% for subject 108 and -25.5% for subject 207.

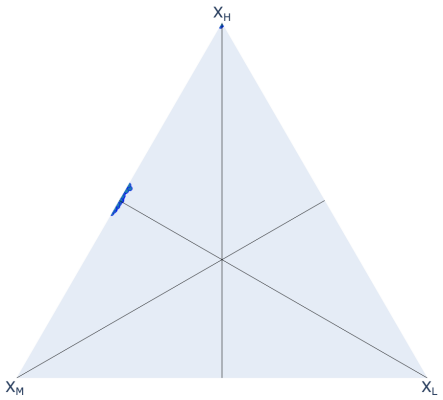
The behavior, however, is quite regular and nearly deterministic; hence, machine learning models exhibit 81.2% and 83.9% completeness. Subjects 108 and 207 exhibits a simple two-part procedural rule. First, if the most expensive account is sufficiently expensive, it is not considered, and as such no Arrow securities are purchased for the corresponding state of the world. Second, the amount of Arrow securities purchased in the remaining states are equalized to remove uncertainty conditional the state(s) being considered. This type of behavior is consistent with rule-based rationality discussed in Choi et al. (2006) and Halevy and Mayraz (2024). In particular, it is consistent with the λ-ratio similarity of Rubinstein (1988). In two dimensions, this is equivalent to disappointment aversion with risk neutrality. However, in three dimensions, the behavior generates a choice rule that is inconsistent with both DA and EUT.

The discrepancy is particularly stark for subject 207. Assuming $p_x < p_y < p_z$, the prediction from risk neutrality always estimates $\mathbf{d} = (1, 0)$, whereas the procedural rule of subject ID 207 will choose $\mathbf{d} = (0.5, 0.5)$, which generates an error of 0.5. Since the uniform random naive mapping expected error is 1/6, completeness is negative.

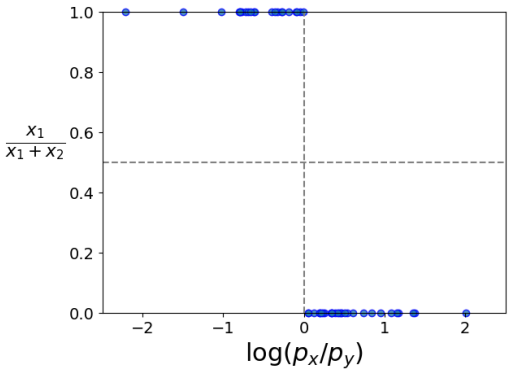
Figure 7: Demand of subjects 108 and 207 for 2D and 3D.



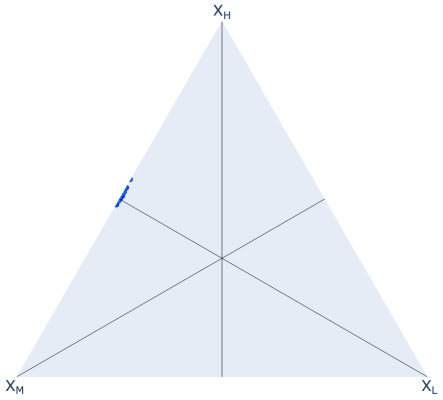
(a) ID 108 2D



(b) ID 108 3D



(c) ID 207 2D



(d) ID 207 3D

The left-hand column shows the relative demand in 2D for the two subjects. The x-axis corresponds to the log price ratio and the y-axis corresponds to the relative demand **d** for account x. The right-hand column shows choices plotted in a rank-ordered budget set, with the most expensive account in the bottom right, the least expensive account in the top, and the final account in the bottom left.

5 Conclusion

Overall, we find strong transferability of economic attitudes across budgetary choice environments. EUT, despite its single-parameter structure, demonstrates strong predictive consistency when transferring demand information between the two-state and three-state environments, with minimal loss in performance relative to within-domain predictions. This consistency underscores EUT's robustness in extrapolating across risk domains, contrasting with DA and machine learning models, which do not achieve substantial improvements or may even underperform when applied out-of-domain. Overall, these findings highlight the utility of economic models in extrapolating behavior across domains and suggest that their theoretical simplicity can be advantageous in contexts where domain shifts occur.

References

Agranov, M. and P. Ortoleva (2017): "Stochastic choice and preferences for randomization," *Journal of Political Economy*, 125, 40–68.

Allais, M. (1953): "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine," *Econometrica*, 21, 503–546.

Andrews, I., D. Fudenberg, L. Lei, A. Liang, and C. Wu (2023): "The Transfer Performance of Economic Models," *arXiv preprint arXiv:2202.04796*.

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

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

Choi, S., R. Fisman, D. Gale, and S. Kariv (2006): "Substantive and procedural rationality in decisions under uncertainty," .

Choi, S., R. Fisman, D. M. Gale, and S. Kariv (2007): "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.

Cox, J. C., V. Sadiraj, and U. Schmidt (2015): "Paradoxes and mechanisms for choice under risk," *Experimental Economics*, 18, 215–250.

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

Ellis, K., S. Kariv, and E. Ozbay (2023): "What can the demand analyst learn from machine learning?" *Working Paper*.

Fehr, E., T. Epper, and J. Senn (2023): "The fundamental properties, stability and predictive power of distributional preferences," .

Friedman, D., S. Habib, D. James, and B. Williams (2022): "Varieties of risk preference elicitation," *Games and Economic Behavior*, 133, 58–76.

Fudenberg, D., W. Gao, and A. Liang (2023): "How flexible is that functional form? quantifying the restrictiveness of theories," *Review of Economics and Statistics*, 1–50.

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, 4112–4141.

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

Gul, F. (1991): "A theory of disappointment aversion," *Econometrica: Journal of the Econometric Society*, 667–686.

Halevy, Y. and G. Mayraz (2024): "Identifying rule-based rationality," *Review of Economics and Statistics*, 106, 1369–1380.

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

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

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

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

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

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.

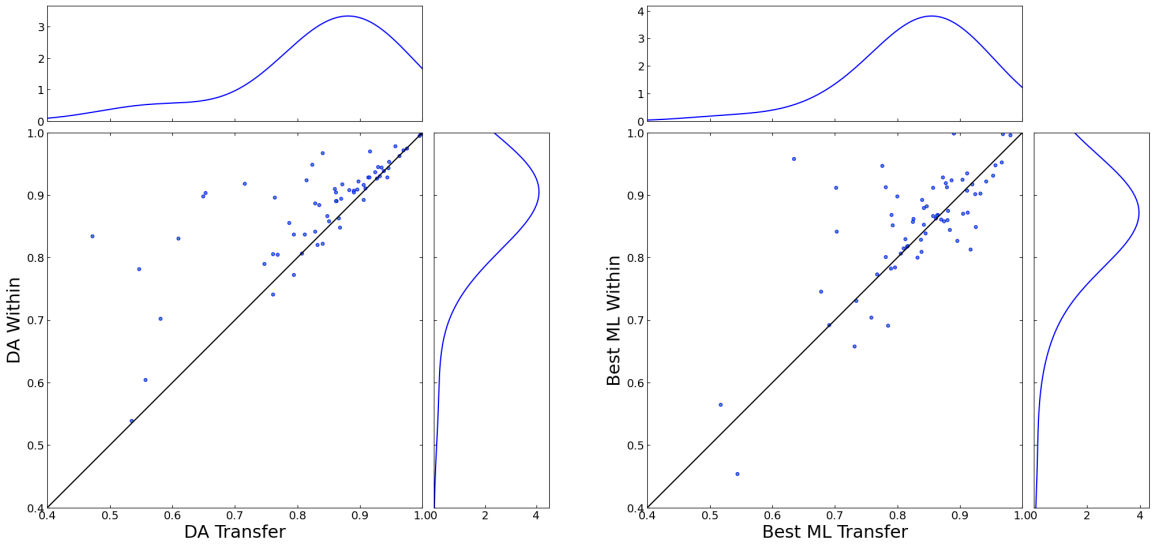
Peysakhovich, A. and J. Naecker (2017): "Using Methods From Machine Learning to Evaluate Behavioral Models of Choice Under Risk and Ambiguity," *Journal of Economic Behavior & Organization*, 133, 373–384.

Plonsky, O., R. Apel, E. Ert, M. Tennenholtz, D. Bourgin, J. C. Peterson, D. Reichman, T. L. Griffiths, S. J. Russell, E. C. Carter, et al. (2019): "Predicting human decisions with behavioral theories and machine learning," *arXiv preprint arXiv:1904.06866*.

Rubinstein, A. (1988): "Similarity and decision-making under risk (Is there a utility theory resolution to the Allais paradox?)," *Journal of economic theory*, 46, 145–153.

Schuler, A., Y. Li, and M. van der Laan (2023): "Lassoed Tree Boosting," *arXiv preprint arXiv:2205.10697*.

A Appendix



(a) Disappointment Aversion (b) Best ML

Figure A.1: Transfer completeness of relative to within-sample completeness.

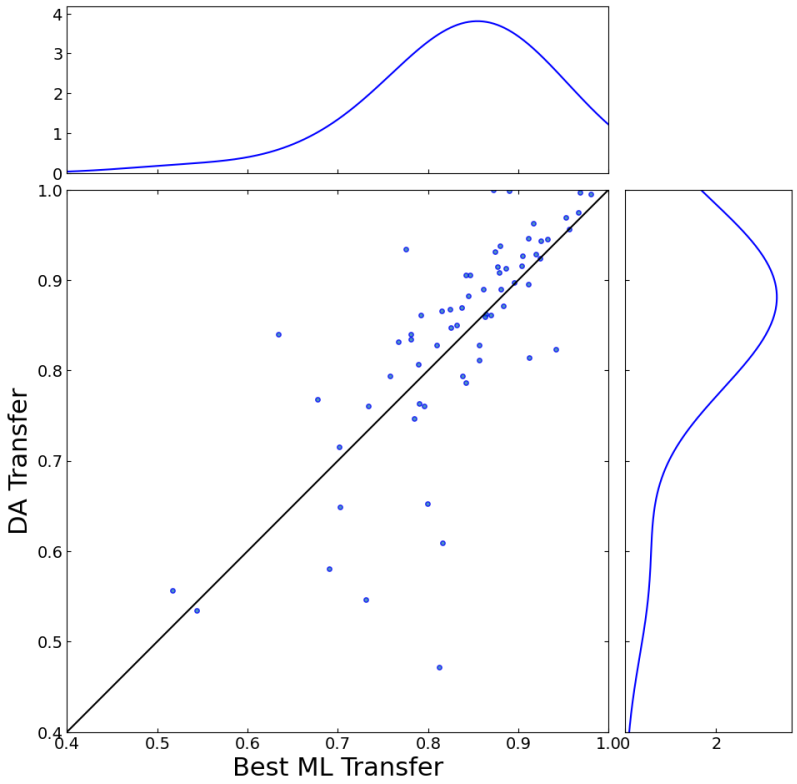


Figure A.2: Relative transfer completeness of disappointment aversion against Best ML