Predicting human decisions with behavioral theories and machine learning

Ori Plonsky[*]
Technion - Israel Institute of Technology
Haifa, Israel
plonsky@technion.ac.il

Reut Apel
Technion – Israel Institute of Technology
Haifa, Israel
reutapel88@gmail.com

Eyal Ert
The Hebrew University of Jerusalem
Jerusalem, Israel
Eyal.Ert@mail.huji.ac.il

Moshe Tennenholtz
Technion – Israel Institute of Technology
Haifa, Israel
moshet@technion.ac.il

David Bourgin
Princeton University
Princeton, NJ
ddbourgin@gmail.com

Joshua C. Peterson
Princeton University
Princeton, NJ
peterson.c.joshua@gmail.com

Daniel Reichman
Princeton University
Princeton, NJ
daniel.reichman@gmail.com

Thomas L. Griffiths
Princeton University
Princeton, NJ
tomg@princeton.edu

Stuart J. Russell
University of California, Berkeley
Berkeley, CA
russell@berkeley.edu

Even C. Carter
United States Army Research Laboratory
Adelphi, MD
evan.c.carter@gmail.com

James F. Cavanagh
The University of New Mexico
Albuquerque, NM
jcavanagh@unm.edu

Ido Erev
Technion – Israel Institute of Technology
Haifa, Israel
erevido@gmail.com

Abstract

Accurately predicting human decision-making under risk and uncertainty is a long-standing challenge in behavioral science and AI. We introduce BEAST Gradient Boosting (BEAST-GB), a hybrid model integrating behavioral insights derived from a behavioral model, BEAST, as features in a machine learning algorithm. BEAST-GB won CPC18, an open choice prediction competition, and outperforms deep learning models on large datasets. It demonstrates strong predictive accuracy and generalization across experimental contexts, highlighting the value of integrating domain-specific behavioral theories with machine learning to enhance prediction of human choices.

[*]An extended version of this paper is available at https://arxiv.org/abs/1904.06866v2

NeurIPS 2024 Workshop on Behavioral Machine Learning.

1 Introduction

Predicting human decision-making under risk and uncertainty is a longstanding challenge in economics, psychology, and artificial intelligence. Despite decades of effort [3, 10, 16, 17, 9], no model accurately describes and predicts choices under risk and uncertainty, even for the most basic stylized tasks, like choice between lotteries (probabilistic options with varying payoffs). Traditional behavioral models aim to describe and explain human choice but often struggle with generalization across contexts, fail to capture the full range of human behavior, and rarely achieve high predictive accuracy [7]. Conversely, pure machine learning (ML) models optimized for prediction accuracy often miss the domain-specific insights essential for capturing human choice behavior effectively and efficiently [14, 12].

This paper introduces BEAST Gradient Boosting (BEAST-GB), a novel hybrid model that combines the predictive power of ML with insights from the behavioral model BEAST (Best Estimate and Sampling Tools). We demonstrate BEAST-GB's efficacy using three large datasets of human choice between lotteries: a recent choice prediction competition (CPC18, [13]), the largest publicly available dataset of human choice under risk and uncertainty [4], and an ensemble of published data used to compare predictions of dozens of classical behavioral models [9]. The results show that BEAST-GB won the prediction competition and outperforms both deep learning models trained on large datasets and existing behavioral models. Importantly, BEAST-GB exhibits strong domain generalization [18], making it effective in novel and unseen contexts. Our findings highlight that integrating behavioral theories with ML models can yield better predictions than using either approach independently.

2 Methods

2.1 BEAST-GB Model Description

BEAST-GB is a hybrid model that combines BEAST, a promising behavioral model of decision-making [7], with Gradient Boosting (XGBoost[5]). The model predicts human decisions in tasks involving choices between lotteries, which form the foundation of rational economic theory [15, 17] and the analysis of deviations from rationality [10, 16]. Integration of the behavioral theory in BEAST-GB is done using engineered features that are derived from BEAST's core assumptions. Notably, these assumptions are very different than those made by mainstream decision models. For example, while mainstream models like prospect theory [10], assume stable preferences, BEAST incorporates mental sampling, acknowledging potential biases and fluctuating decision-making processes alongside sensitivity to expected values. Specifically, BEAST assumes decision-makers use a small mental sample of possible outcomes from available options, with a potentially biased sampling process. The mental sampling process reflects four behavioral tendencies: choosing options that minimize immediate regret, choosing options that maximize the worst outcome, choosing options that maximize the probability of the best possible payoff sign, and choosing options that are expected to lead to the best outcome had all of the outcomes been equally likely.

The features used in BEAST-GB are defined by Plonsky et al. [14] and presented in Table 1. In addition to the set of raw features that describe the available lotteries (referred to as *objective* features), BEAST-GB leverages three sets of *behavioral* features:

- **Naïve** features: Simple properties like differences in expected values or standard deviations.
- **Psychological insight** features: Capturing deeper behavioral mechanisms from BEAST, such as sensitivity to the probability that one option provides a better outcome than another in a single draw, corresponding with BEAST's assumption that people choose options that minimize immediate regret.
- **Behavioral foresight** feature: The direct prediction of BEAST for each choice task.

2.2 Datasets

We evaluate BEAST-GB on three datasets that assess human decision-making under risk and uncertainty (see A.3 for more details).

CPC18 An open choice prediction competition (CPC) dataset containing 270 binary choice tasks between lotteries (see Figure 1). The competition's training set included 210 tasks, while 60 tasks served as the held-out test set. Choice tasks were randomly selected from a predefined space of decisions under risk, ambiguity, and experience (the "Objective" features in Table 1), and cover classical behavioral phenomena like the Allais [2] and Ellsberg [6] paradoxes. The dataset consists of 694,500 decisions made by 926 subjects across 25 trials each, first without feedback and then with feedback on the outcomes of previous choices.

Choices13k The largest publicly available dataset for human choice under risk, Choices13k includes 9,831 binary choice tasks, similar to CPC18 (i.e, with the same objective features), where Amazon Mechanical Turk users made five repeated choices between lotteries with feedback. Introduced to explore ML models' predictive power in larger datasets [4, 12], Choices13k allows comparison of BEAST-GB with state-of-the-art pure ML models.

HAB22 An ensemble dataset containing 1,565 one-shot choice tasks from various published studies, recently assembled for a large-scale comparison of over 50 behavioral models of risky choice [9]. The dataset includes tasks that differ from those in CPC18 and Choices13k, featuring large differences in expected values and no feedback, and allows testing model robustness and generalization across its 15 different experimental contexts.

2.3 Evaluation

Following previous works [7, 12], we focus on the models' ability to predict the aggregate behavior of decision-makers. Hence, the models aim to predict choice rates—the proportion of participants choosing each option in a new task. We assess model performance using two key metrics: **Mean Squared Error (MSE)**, which quantifies the difference between the predicted and observed choice rates, measuring the model's accuracy, and a **completeness score** (see A.1), representing the proportion of predictable variation in the data that the model captures[8]:

$$\text{Completeness} = \frac{\text{MSE}_{\text{random}} - \text{MSE}_{\text{model}}}{\text{MSE}_{\text{random}} - \text{MSE}_{\text{irreducible}}}$$

where $\text{MSE}_{\text{random}}$ is the MSE for random guessing and $\text{MSE}_{\text{irreducible}}$ is the irreducible sampling error. This score evaluates how well the model captures systematic behavioral patterns beyond random fluctuations.

In CPC18, which includes a hidden test set, the models are compared based only on their predictive accuracy in the (single) test set. In Choices13k and HAB22, models are evaluated based on 50 repetitions of a 90-10 split of the choice tasks to train and test sets, with the reported metrics the average of the performance in the 50 test sets.

2.4 Feature Importance Analysis

To assess which features contribute most to the predictive performance of BEAST-GB, we conduct two types of feature importance analysis:

Feature Removal: We systematically remove sets of features (naïve, psychological, or foresight) and retrain the model to evaluate how prediction accuracy changes. This process helps determine the relative importance of each feature set.

SHAP Values: We compute SHAP (Shapley Additive Explanations) values [11] to quantify the contribution of individual features to each prediction. SHAP values provide insight into how different behavioral features influence the model's predictions.

3 Results

3.1 Predictive accuracy

CPC18 Sixty-nine researchers from 34 institutions across 16 countries registered to submit predictions to the CPC, with 20 models (including BEAST-GB) submitted on time. BEAST-GB won the competition, achieving the lowest MSE (0.0056) on the held-out test set. Its completeness score

was 93%, indicating that it captures nearly all of the predictable variability in the data. Notably, submissions of pure ML approaches performed significantly worse.

Choices13k BEAST-GB set a new state-of-the-art in predictive accuracy for this dataset, with an MSE of 0.0084 and a completeness score of 96%. This accuracy is 25% better than the most expressive neural network model previously reported [12], which did not leverage behavioral theory. Additionally, we tested the sample efficiency by training the models on increasing proportions of the train data. As shown in Figure 2, BEAST-GB demonstrated significant improvements in learning speed: Trained on just 2% of the data (176 tasks), it surpassed the accuracy of a neural network trained on the entire dataset (̃9000 tasks).

HAB22 BEAST-GB significantly outperformed all 50+ behavioral models previously tested on the HAB22 dataset, including various versions of prospect theory, achieving a completeness score of 95% (see Figure 3). This suggests that BEAST-GB captures systematic patterns in human behavior even in one-shot tasks without feedback, underscoring its robustness. Interestingly, despite BEAST's relatively poor performance on this dataset (completeness score of 36%), BEAST-GB's integration with machine learning enables it to excel, highlighting the value of ML in enhancing less flexible behavioral models.

3.2 Feature importance analysis

Feature importance analysis consistently highlighted the importance of the behavioral foresight feature (BEAST's prediction). Specifically, In SHAP value analysis, the foresight feature consistently had the largest SHAP value across all datasets, and in the feature removal analysis, removing this feature usually led to the largest reduction in accuracy. This highlights the critical role of BEAST's original predictions, even in cases where its point predictions are poor, as observed in the HAB22 dataset. This may imply that BEAST captures key behavioral patterns even though it is biased in some contexts, and highlights how the use of ML can improve predictions of good but inflexible behavioral models.

Interestingly, in the Choices13k dataset, removing the foresight feature early in training severely impaired prediction accuracy, but its impact diminished as more data was used. Eventually, the model performed nearly as well without it, suggesting that with sufficient data, the ML component of BEAST-GB can learn the proper integration of the psychological insights underlying BEAST without direct access to BEAST's predictions. However, removal of the psychological insight features themselves resulted in a significant decline in performance, even with large training data, and the worst performance occurred when both the psychological insight and the behavioral foresight features were removed, underscoring the continued importance of behavioral theory in prediction tasks, even with large datasets.

3.3 Domain generalization

While BEAST-GB excels in making accurate predictions within specific experimental contexts, we also examined its ability to generalize across different contexts. Using the HAB22 dataset, which contains 15 distinct experimental contexts, we trained BEAST-GB on data from 14 contexts and tested its performance on the 15th. Remarkably, BEAST-GB achieved a completeness score of 85%, predicting well even without access to data from the unseen context.

Further analysis showed that BEAST-GB generalizes better across contexts than observed participant behavior. When predicting choice rates of known tasks across different contexts, BEAST-GB reached a completeness score of 92%, significantly outperforming a model that assumes that the behavior of one sample of participants in an experimental context generalizes to another sample in another context (in an identical choice task). These results (Figure 4) suggest that BEAST-GB captures generalizable decision-making patterns that transcend individual contexts.

4 Discussion

BEAST-GB demonstrates the strength of hybrid models by successfully integrating behavioral theory with machine learning, offering a novel approach to predicting human decision-making under risk

and uncertainty. Its ability to incorporate BEAST, an interpretable model of human choice, into a powerful machine learning framework highlights the value of combining qualitative insights from behavioral science with the quantitative precision of AI systems.

A key strength of BEAST-GB is that it derives its predictive power from BEAST's clear, interpretable assumptions, such as mental sampling and sensitivity to expected values. This makes the model highly explainable compared to typical machine learning models, which often function as black boxes. BEAST-GB's predictions are highly correlated with those of the behavioral model BEAST, which ensures that the behavioral foundation is preserved while allowing the model to overcome BEAST's limitations, such as bias in specific contexts. SHAP analyses further support this, showing that BEAST's predictions significantly contribute to the model's overall performance, even in domains where BEAST alone performs poorly.

This hybrid approach also generalizes well across domains, as evidenced by BEAST-GB's performance on unseen experimental contexts within the HAB22 dataset. Not only does the model excel in its trained contexts, but it also outperforms direct empirical generalizations of human behavior in new domains, indicating its robustness and broad applicability. Further, because BEAST-GB is heavily based on an interpretable psychological model that assumes similar behavioral tendencies regardless of the exact problem structure, it should be straightforward to generalize BEAST-GB to other tasks of decisions under risk and uncertainty. For example, to predict behavior in multiple (rather then binary) choice tasks, one only needs to tweak the Psychological Insight features of BEAST-GB so they would capture the same behavioral tendencies (like choosing options that minimize immediate regret) but given more choice options.

The development of BEAST-GB has implications beyond choice prediction under risk. We replicated the process by which this hybrid model was developed (transformation of a strong behavioral model to features ingrained in ML) in a different domain—human choice in Market Entry Games—and achieved similar success. This demonstrates that this approach has the potential for broader applicability in other areas of human behavior.

Furthermore, the scalability of BEAST-GB is a notable advantage. BEAST-GB simplifies the use of BEAST by incorporating its features without the need for retraining, making it more computationally efficient. This efficiency is especially important in large datasets like Choices13k, where BEAST-GB learns much faster than purely data-driven models and requires fewer training examples to achieve high accuracy. This suggests that behavioral insights play a crucial role in making machine learning models both faster and more generalizable.

In conclusion, BEAST-GB illustrates how integrating behavioral theories into machine learning models can enhance predictive accuracy, interpretability, and generalization. As the workshop explores interdisciplinary approaches to behavioral modeling, BEAST-GB serves as a compelling case for the potential of hybrid models to bridge the gap between behavioral science and AI.

References

[1] Or David Agassi and Ori Plonsky. The Importance of Non-analytic Models in Decision Making Research: An Empirical Analysis using BEAST. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

[2] Maurice Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, 21(4):503–546, 1953.

[3] Daniel Bernoulli. Exposition of a new theory on the measurement of risk (original 1738). *Econometrica*, 22(1):23–36, 1954.

[4] David D Bourgin, Joshua C Peterson, Daniel Reichman, Stuart J Russell, and Thomas L Griffiths. Cognitive model priors for predicting human decisions. In *International conference on machine learning*, pages 5133–5141. PMLR, 2019.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[6] Daniel Ellsberg. Risk, ambiguity, and the Savage axioms. *The quarterly journal of economics*, 75(4):643–669, 1961.

[7] Ido Erev, Eyal Ert, Ori Plonsky, Doron Cohen, and Oded Cohen. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, 124(4):369–409, 2017.

[8] Drew Fudenberg, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. Measuring the completeness of economic models. *Journal of Political Economy*, 130(4):956–990, 2022.

[9] Lisheng He, Pantelis P Analytis, and Sudeep Bhatia. The wisdom of model crowds. *Management Science*, 68(5):3635–3659, 2022.

[10] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

[11] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[12] Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.

[13] Ori Plonsky, Reut Apel, Ido Erev, Eyal Ert, and Moshe Tennenholtz. When and how can social scientists add value to data scientists? A choice prediction competition for human decision making. Technical report, 2018.

[14] Ori Plonsky, Ido Erev, Tamir Hazan, and Moshe Tennenholtz. Psychological forest: Predicting human behavior. In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017.

[15] Leonard J Savage. *The foundations of statistics*. John Wiley I& Sons, Oxford, 1954.

[16] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, oct 1992.

[17] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton university press, Princeton, 2 edition, 1947.

[18] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

A supplemental material - Additional Methods

A.1 Completeness Score Calculation

Completeness is defined as:

$$\text{Completeness} = \frac{\text{MSE}_{\text{random}} - \text{MSE}_{\text{model}}}{\text{MSE}_{\text{random}} - \text{MSE}_{\text{irreducible}}}$$

where $\text{MSE}_{\text{random}}$ is the MSE of random guessing (as defined in Fudenberg et al.[8]), $\text{MSE}_{\text{model}}$ is the MSE of the model in question, and $\text{MSE}_{\text{irreducible}}$ represents the irreducible error, or the part of the total error presumed to be unpredictable.

To compute $\text{MSE}_{\text{irreducible}}$, we estimate the expected MSE of a perfect hypothetical model that accurately predicts the population choice rate in a task. The observed error of the perfect theoretical model in task i is the sampling error. The computed $\text{MSE}_{\text{irreducible}}$ can be expanded as follows:

$$\text{MSE}_{\text{irreducible}} = \frac{1}{N}\sum_{i=1}^{N}(\mu_i - \bar{x}_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(\mu_i - x_i)^2$$

where μ_i is the true population choice rate for task i and x_i is the observed sample choice rate.

The expected value of $\text{MSE}_{\text{irreducible}}$ is therefore:

$$E(\text{MSE}_{\text{irreducible}}) = \frac{1}{N}\sum_{i=1}^{N} E\left((\mu_i - \bar{x}_i)^2\right) = \frac{1}{N}\sum_{i=1}^{N} \text{Var}(x_i) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{S_i^2}{n_i}$$

where S_i^2 is the sample variance for task i, and n_i is the sample size for task i. Therefore, our estimate for $\text{MSE}_{\text{irreducible}}$ is the average of the squared standard errors.

A.2 BEAST-GB training and implementation.

BEAST-GB is an XGB algorithm that uses the features detailed in Table 1. We implemented the following pipeline to train BEAST-GB on each choice dataset. First, we generated the features for each choice task. This notably includes generating the choice rate prediction of the original BEAST model for that choice task (without refitting BEAST to the new data). Second, we coded categorical features to numeric using dummy coding. Third, because in particular datasets some features may turn out completely constant and/or duplicates of other features, we removed such features from the data. Fourth, we randomly split the data to a train and a held-out test set (unless the data was already organically split, like in CPC18). Fifth, we standardized all features by subtracting their mean and dividing by their standard deviation in the train set. Sixth, we tuned the algorithm's hyperparameters using five repetitions of 5-fold cross validation implemented on the train set. Finally, we trained the algorithm on the full train set with the chosen hyperparameters and generated its predictions for the held-out test set.

A.3 Dataset details

Similar to the paradigm used in CPC15 [7], the experimental paradigm in CPC18 involved binary choice under risk, under ambiguity, and from experience. As seen in Figure 1, decision-makers are faced with descriptions of two lotteries (Option A and Option B) and are asked to choose between them repeatedly for 25 trials. In the first five trials, they do not get any feedback, but as of the 6th trial they get full feedback concerning the outcomes generated by each option (both the obtained and the forgone payoffs are revealed). Choice options in CPC18 may include up to 10 outcomes, may involve ambiguity (i.e., probabilities of potential outcomes are not revealed to the decision maker), and may be correlated between them.

The Choices13k dataset was originally presented by Bourgin et al.[4] and includes 13,006 binary choice tasks. Tasks were generated by the task generation algorithm used in CPC15[7] and are therefore all members of the same space used in CPC18 that extends it. The data includes, for each choice task, the proportion of times in which participants chose Option B. As in Peterson et al.[12], we removed from the dataset tasks in which one of the options was ambiguous and tasks in which participants did not receive any feedback, resulting with a dataset containing 9,831 risky choice tasks in which participants made five consecutive choices with full feedback after each choice

HAB22 includes data assembled by He et al[9] from 15 different experimental contexts published in seven distinct papers by various researchers. In each experimental context, participants made multiple one-shot choices between binary lotteries with up to two outcomes without feedback. Hence, the experimental task here is different than those used in CPC18 and Choice13k. Moreover, some choice tasks in this dataset are very different than those used in the other two datasets of choice under risk and uncertainty we use in this paper. Specifically, the difference between the EVs of the lotteries in some choice tasks here are especially large. In total, the HAB22 data includes 1565 choice tasks, although some of these are identical but were run in different experimental contexts and are thus treated as distinct.

Originally, He et al [9] used four additional experimental contexts in their analyses. However, the data in these contexts is not usable for our purposes[1]. In three contexts, the data they used to train their behavioral models included errors (specifically, the actual choice data was wrongly linked with the identity of the choice tasks participants observed) and in a fourth, participants faced the same choice tasks more than once, leading the same choice task to be included both in the train and test

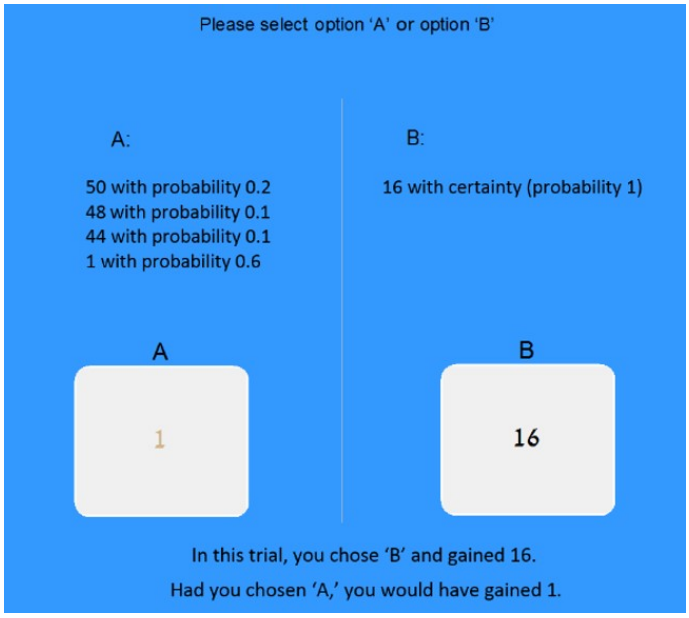

Figure 1: A decision-making task used in CPC18. Human decision makers choose between A and B repeatedly for 25 trials, and get full feedback after each choice starting trial 6. The participant here chose Option B.

set of the behavioral models. Hence, we could not properly compare BEAST-GB to the behavioral models in these contexts and thus chose to drop them.

B supplemental material - Additional Results

The following figures present the main results of the comparative analyses in Choices13k and HAB22.

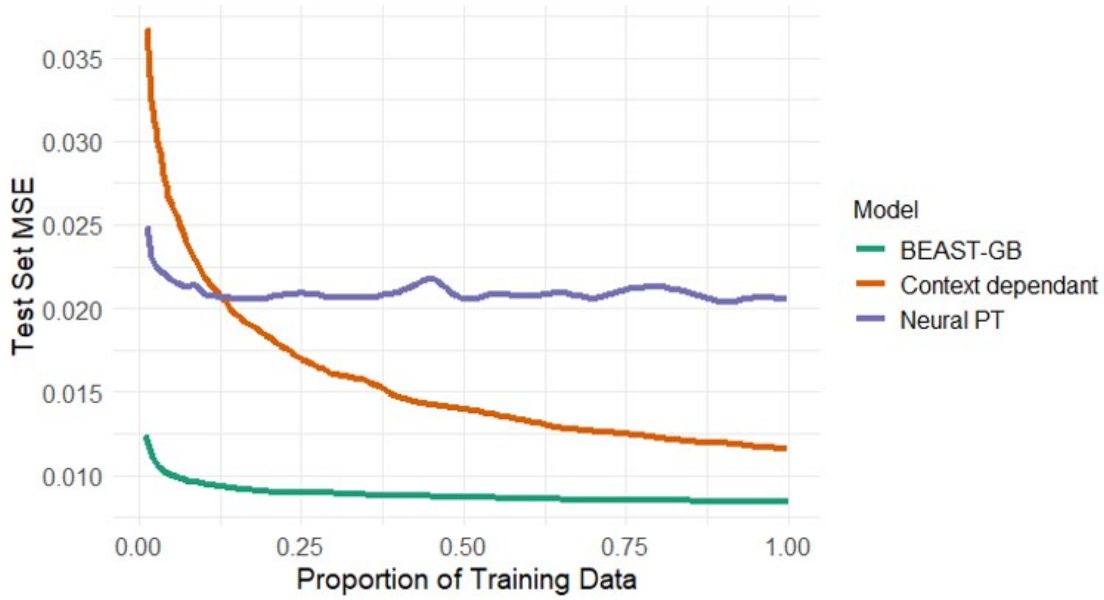

Figure 2: Test set performance on Choices13k data. Data was split to 90% training and 10% held-out test data, and models were trained on fixed and increasing proportions of the training data. This process was repeated 50 times, and results reflect the average test set MSE over these 50 splits. Neural PT (Neural Prospect Theory) and Context dependent performance is taken directly from Peterson et al. (2021).

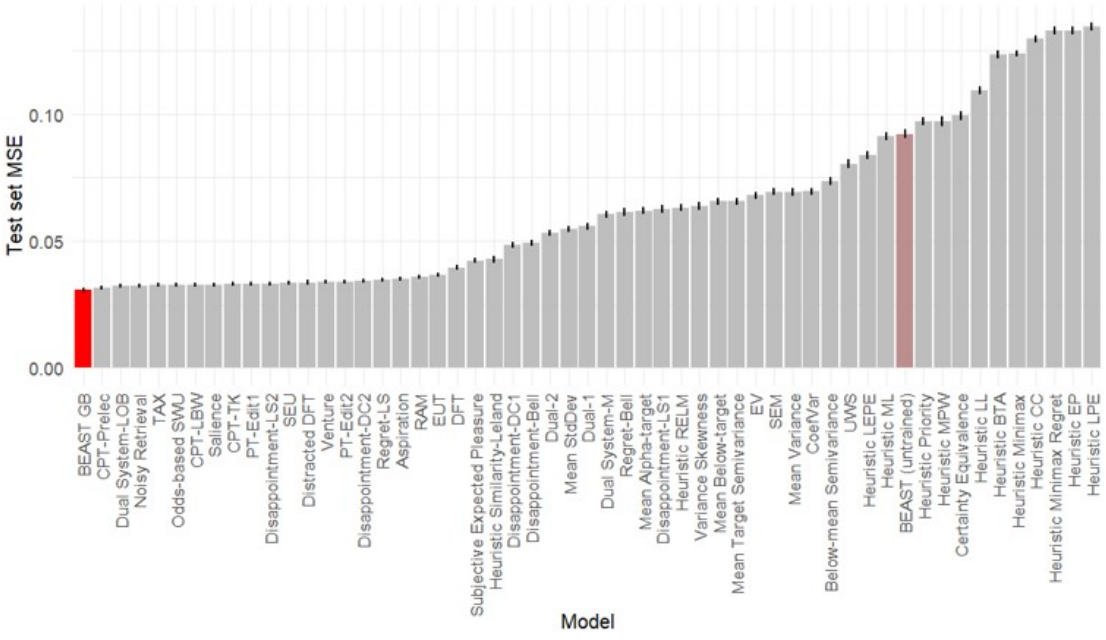

Figure 3: Test set performance on HAB22 data. Performance is evaluated based on 10-fold cross validation on choice tasks, and 5-fold cross validation on participants in experimental contexts. That is, models predict choice rates of new participants in new tasks (see Methods). Error bars represent ±1 SE for the mean over the 50 test sets. Models, except BEAST and BEAST-GB, were evaluated by He et al., 2022 and details on the models can be found there.

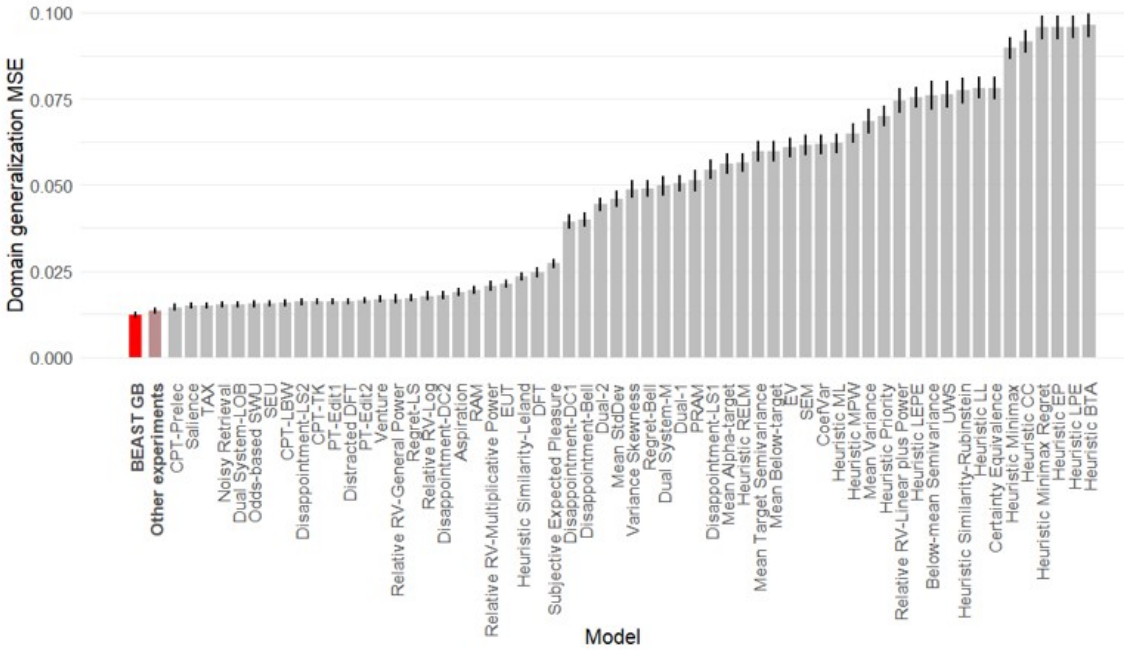

Figure 4: Predictive accuracy in domain generalization task of predicting behavior in the 762 instances where a choice task that appears in the test dataset also appeared in one or more of the train datasets. Training data always includes 14 experimental contexts to predict the 15th context. Behavioral models' predictions are set to be the average training prediction (i.e., best fit) in the target task across all subjects in the training data. "Other experiments" prediction is the average observed behavior across all subjects in the training data in the target task. Error bars represent ±1 SE for the mean over the 762 prediction errors. Models, except BEAST and BEAST-GB, were evaluated by He et al., 2022 and details on the models can be found there.

Table 1: Features used in BEAST-GB

Name	Description
Objective	
Ha	H_A - High payoff in Option A. When Option A has multiple outcomes, H_A is the EV of the lottery in Option A.
pHa	pH_A - Probability of H_A.
La	L_A - Low payoff in Option A.
LotShapeA	Shape of the distribution of the lottery in Option A.
LotNumA	Number of outcomes in distribution of the lottery in Option A. When Option A does not have multiple outcomes, $LotNumA = 1$.
Hb	H_B - High payoff in Option B. When Option B has multiple outcomes, H_B is the EV of the lottery in Option B.
pHb	pH_B - Probability of H_B.
Lb	L_B - Low payoff in Option B.
LotShapeB	Shape of the distribution of the lottery in Option B.
LotNumB	Number of outcomes in distribution of the lottery in Option B. When Option B does not have multiple outcomes, $LotNumB = 1$.
Amb	Indicator for an ambiguous choice task (1 if True, 0 otherwise).
Corr	Sign of correlation between generated payoffs in the two options (-1, 0, or 1).
block	The block number in repeated choice tasks (each block corresponds to 5 trials).
Feedback	Indicator for block with feedback (1 if True, 0 otherwise).
Dataset	Dataset from which task is taken.
Naive	
diffEVs	Difference between the payoff EV of Option B and the payoff EV of Option A.
diffSDs	Difference between the payoff SD of Option B and the payoff SD of Option A.
diffMins[a]	Difference between the minimal payoff of B and the minimal payoff of A.
diffMaxs	Difference between the maximal payoff of B and the maximal payoff of A.
Psychological Insight	
diffBEV0	Difference between the "best estimate" of the EVs as per BEAST, before feedback.
diffBEVfb	Difference between the "best estimate" of the EVs as per BEAST, after feedback.
pBbet_Unbiased1	Difference between the probability that Option B provides better payoff than Option A, as estimated by BEAST before feedback.
diffUV	Difference between the EV of Option B when all its outcomes are transformed to be equally likely and the EV of Option A when all its outcomes are equally likely.
pBbet_Uniform	Difference between the probability that Option B provides better payoff than Option A, when both options are transformed so that their outcomes are equally likely.
RatioMin	Ratio between the smaller and the higher minimal outcomes of the two options.
SignMax	The sign of the maximal possible payoff in the task (-1, 0, or 1).
diffSignEV	Difference between the EV of Option B when all outcomes are sign-transformed and the EV of Option A when all outcomes are sign-transformed.
pBbet_Sign1	Difference between the probability that Option B provides better payoff than Option A, as estimated by BEAST before feedback and after all payoffs are sign transformed.
pBbet_SignFB	Difference between the probability that Option B provides better payoff than Option A, as estimated by BEAST after feedback and after all payoffs are sign transformed.
Dom	Trinary indicator for the option that stochastically dominates another (1 = B dominates A; -1 = A dominates B; 0 = neither option has dominance).
Behavioral Foresight	
BEASTpred	The quantitative point prediction of BEAST for the choice task (and block).

Notes: This is an exhaustive list of every feature used in this paper as part of BEAST-GB. When run on particular datasets, some features may be completely constant and others may be duplicates of other existing features. In such cases, these features are removed before running the algorithm.
[a] `diffMins` belongs to both the naive and the psychological feature sets.