Title
The Importance of Non-analytic Models in Decision Making Research: An Empirical Analysis using BEAST

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Agassi, Or David
Plonsky, Ori

Publication Date
2023

Peer reviewed

The Importance of Non-analytic Models in Decision Making Research: An Empirical Analysis using BEAST

Or David Agassi (odagassi@gmail.com), Ori Plonsky (plonsky@technion.ac.il)
The Faculty of Data and Decision Sciences, Technion – Israel Institute of Technology, Haifa 3200003, Israel

Abstract

Decision-making models hold a vital role in the field of cognitive science, serving as a means of describing and predicting human behavior. While classical models with similar assumptions are frequently favored, there is no guarantee they provide the best accounts of behavior. Here, we evaluate BEAST, a model that has demonstrated extraordinary predictive capabilities in diverse settings, but was excluded from a recent large-scale comparison of models because it cannot be analytically estimated. Our evaluation of the model's performance on a large collection of experiments of decisions under risk shows it provides excellent predictions in some domains. We further show how BEAST can be adapted to increase its predictive power in contextualized settings. Our results highlight the importance of a more inclusive approach toward models that may be difficult to analytically estimate to deepen our understanding of the psychological mechanisms underlying human decision making behavior.

Keywords: Decision making; Computational modeling

Introduction

Decision-making under risk and uncertainty is a fundamental aspect of human cognition. Understanding how individuals approach and make decisions in the face of risk has been the focus of much research in the field of cognitive science. This is motivated by the recognition that individuals often make decisions that deviate from the predictions of classical normative models of decision-making under risk, such as expected utility theory. In light of this, researchers have sought to develop and test alternative models that can better describe and predict human decision-making behavior.

Through the years, numerous models have been developed to capture human decision-making, ranging from simple mathematical models to more complex process models incorporating various psychological mechanisms underlying risky choices. The proliferation of various models raises a challenge to understand which mechanisms and models are the most applicable for describing and predicting human decision-making behavior in different environments. To overcome this challenge, systematic evaluation and comparison of models can be very useful.

One effective way to perform such systematic evaluations is to compare models based on their predictive accuracy in large sets of human choice problems, preferably originating from different participants participating in different decision making experiments. The methodology of comparing models based on prediction accuracy on common data draws from a large literature in computer and data science, facilitates comparison between models with different number of parameters, and increases the chances that diverse models will be developed (Plonsky & Erev, 2021).

In a recent impressive study, He, Analytis, and Bhatia (2022) performed a large-scale comparison of dozens of models of decision-making under risk. They grouped 19 different published datasets from different papers with more than 1800 different choice problems. In each problem, participants were asked to make a one-shot decision between two fully described gambles. Each gamble included up to two possible outcomes with known probabilities and either involved only potential gains (i.e., problems from the "gain domain"), or both potential gains and losses (i.e., problems from the "mixed domain"). This paradigm of choice between gambles has been a prevalent research tool in behavioral economics since its inception, enabling researchers to gain valuable insights into human preferences and attitudes in a wide range of decision-making contexts (Allais, 1953; Ellsberg, 1961, Erev et al., 2017, Ert & Erev, 2013; Kahneman & Tversky, 1979; Stewart, Reimers, & Harris, 2015). In their study, He et al. compared between 58 published models of risky choice to examine which of these offers the best accounts of choice behavior. The results showed that crowd models, which aggregate the predictions of multiple models, outperform individual models. They concluded that different existing models function as complementary rather than competing accounts of behavior.

Despite the large collection of models that He et al. (2022) compared in their study; some extant models were excluded from their analysis. In particular, they chose to exclude models that could not be fitted easily using analytical likelihood functions, like those that require running simulations to make predictions. Indeed, fitting such models requires significant efforts. The choice to exclude this type of models may reflect a general practice in the field that tends to focus on models that are amenable to estimation and whose parameters can be easily identified. The focus on such models diminishes modelling effort, allows building directly on previous classical models (like expected utility and prospect theory) and is likely more easily justifiable to reviewers and readers (Plonsky & Erev, 2021). However, there is no a-priori reason to assume that a theoretical "ideal model" of decision-making must necessarily fall within the space of models that are easily estimable. Ignoring non-analytic models may

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

hinder progress and suppress our understanding of human decision making (Bugbee & Gonzalez, 2022).

This potential problem may be particularly concerning as models that utilize simulations to generate predictions—and that are therefore not easily estimable using traditional fitting practices—have a strong track record of accurately predicting behavior (Erev, Ert, & Roth, 2010; Erev et al., 2017; Plonsky et al., 2019). One prominent model in this class is BEAST (Best Estimate and Sampling Tools), a simulation-based model that was designed to predict choice between economic prospects over time (Erev et al., 2017). The model was demonstrated to capture 14 different anomalies in human choice behavior (including Allais', St. Petersburg', and Ellsberg's paradoxes) and was used as a basis for the best performing models in two choice prediction competitions (Erev et al., 2017; Plonsky et al., 2019). Notably, BEAST was originally developed to capture choice in a very wide class of problems that include simple decisions under risk tasks (like those used by He et al., 2022) but also decisions under ambiguity and decisions under risk with repeated feedback. Due to its wide coverage, developers of BEAST, concerned with overfitting issues, introduced several arbitrary implementation assumptions that restrict the model in ways that are not necessarily implied by the underlying theory, but save free parameters.

In this paper, we demonstrate the potential pitfalls of disregarding non-analytic models in the field of decision-making research. To do so, we use BEAST, a model which has demonstrated extraordinary predictive capabilities in decision making competitions but was excluded from the analysis conducted by He et al. (2022) due to its challenging estimation process. We apply similar methods to those used by He et al. and examine the predictive power of BEAST on their data. Our analysis reveals that BEAST achieves strong performance in the mixed gain/loss domain of choice problems, ranking as one of the best performing models for this data. This finding is particularly noteworthy as the model is very different than the most successful models of the original study. For example, it operates primarily through the use of sampling and regret mechanisms that were not used by the other top-performing models in that study. However, BEAST falls short when applied to problems from the gain domain. Further analysis suggests the main reason for this subpar performance is some of the model's strong original arbitrary implementation assumptions that considerably hurt its flexibility. When applied to one-shot decisions under risk, like the datasets in He et al. (2022), BEAST requires less free parameters, and thus it is also possible to relax some of the restrictive implementation assumptions and increase the model's flexibility. We therefore developed Weighted-BEAST (W-BEAST), a flexible modification of BEAST, and applied it to the same data. The results show an immense improvement in performance in the gain domain, without compromising the excellent performance in the mixed domain. The results also clarify when the original assumptions are likely to hurt the model.

We thus make two main contributions. We demonstrate that overlooking non-analytic that are not easily estimable may lead to ignoring strong models of decision making, and we also suggest an improved version of BEAST for the case of decisions under risk that also allows a more contextualized analysis of the model. This highlights the importance of considering a range of models, including those that may be considered more difficult to estimate, as it can add valuable insights into the underlying mechanisms of human behavior.

Method

BEAST

In binary decision under risk problems, the original BEAST model implies option A is preferred over option B if:

$$[EV_A - EV_B] + [ST_A - ST_B] + e > 0$$

where $EV_A - EV_B$ is the advantage of option A over option B based on the expected values (EVs), $ST_A - ST_B$ is the advantage of option A over option B based on mental sampling using sampling tools, and e is a normally distributed error term with a mean 0 and standard deviation σ_i, where i represents an individual (if one option stochastically dominates the other, $e = 0$).

ST is the average of κ_i outcomes that are each generated by using one of four possible sampling tools. Each sampling tool represents a different strategy to mentally draw outcomes. Sampling tool *unbiased* implies simple unbiased draw from the options' described distributions. The remaining three sampling tools imply biased sampling. Sampling tool *uniform* neglects the described probabilities and assumes an equal probability for each outcome (Thorngate, 1980). Sampling tool *contingent pessimism* yields the worst possible outcome (Edwards, 1954) under some lexicographic conditions (Brandstatter, Gigerenzer, & Hertwig, 2006) that depend on some value γ_i. Sampling tool *sign* is highly sensitive to the payoff sign but ignorant of the size of the outcomes (Payne, 2005). BEAST assumes that the probability to use each of the three biased sampling tools is equal, and the probability to use the unbiased tool is $1 - \frac{\beta_i}{\beta_i + 1}$. The individual parameters are drawn from uniform distributions as follows: $\sigma_i \sim U[0, \sigma]$, $\kappa_i \sim \mathrm{U}(1,2,\ldots,\kappa)$, $\gamma_i \sim U[0, \gamma]$, $\beta_i \sim U[0, \beta]$ with $\sigma, \kappa, \gamma, \beta$ as free parameters. For further details about the model and its mechanisms kindly see Erev et al (2017).

Weighted BEAST (W-BEAST)

As explained above, BEAST was designed to capture behavior under diverse conditions, including decisions under risk with multiple outcome gambles, decisions under ambiguity and decisions from experience. To deal with this complexity, its developers made several arbitrary implementation assumptions that restrict the model but save free parameters. Since here we focus on decisions under risk with up to two outcomes, W-BEAST relaxes these arbitrary restrictions.

Sampling Tools. The original BEAST included a very restrictive assumption which prescribed equal probability to use each of the biased sampling tools. However, different datasets, each with different experimental settings, can trigger the preferential usage of certain sampling tools. This is particularly true in datasets involving decision under risk without feedback as used in this study. To capture possible contextual effects, W-BEAST relaxed this restrictive assumption. In our modification, the probability to use each of the biased sampling tools is a free parameter. Specifically, W_{uf}, W_s and W_{cp} represent the probability to use the *uniform*, *sign*, and *contingent pessimism* tools respectively. The probability to use the *unbiased* sampling tool is then simply $W_{ub} = 1\text{-}(W_{uf} + W_s + W_{cp})$.

Expected Value. The original BEAST made an arbitrary assumption of assigning the same weight for the difference between the EVs and for the difference between the average of the mental samples. This assumption precluded the possibility that the significance of the EV may fluctuate depending on the specific context or dataset being utilized. In W-BEAST, we entirely eliminated the use of the EV. Instead, we relied on the fact the many unbiased samples from the true distributions approximate the EV of the gamble. Hence, with sufficiently high value for W_{ub}, which represents the weight of the unbiased sampling tool, and sufficiently large κ_i, which represents the number of samples drawn, the mental sampling process can capture sensitivity to EVs. Of course, small values for these parameters imply low weight to the EV. In this manner, W-BEAST is able to express a wide spectrum of weights to the EV.

Sampling Size. Finally, in W-BEAST, instead of drawing κ_i from a uniform distribution, it is drawn from a geometric distribution with parameter p. That is, the sample size equals $Pr(\kappa_i = k) = (p-1)^{k-1}p$. This change is based on the observation that most participants behave as if they rely on small samples (e.g., Plonsky, Teodorescu, & Erev, 2015), but allows for some participants to rely on large ones.

Estimation and Cross Validation

Fitting the models to the new data requires estimation of the parameters $\sigma, \kappa, \gamma, \beta$ for BEAST and the parameters $p, \sigma, W_{uf}, W_s, W_{cp}$ for W-BEAST. Because the models are simulation-based and do not have a differentiable likelihood function, we calculated the likelihood of each profile of parameters, with each free parameter allowed to get one of several values taken from a closed set of possible values. The sets of values for all the parameters for each dataset are detailed in the supplementary material (SM; see https://osf.io/q45kf).

The estimation used a cross-validation technique similar to that was used for the other 58 models in the original study of He et al (2022). First, we generated predictions for all the problems for each possible profile of parameters. The choice data for each participant was then split into the exact 10 folds of problems as in the original study. In each cross-validation iteration, we chose the profile of parameters that best fits 9 of these folds (representing 90% of the choice data), based on maximum likelihood criterion (Cousineau & Allen, 2015), and then elicited the fitted models' predictions for the held-out fold. This process was repeated 10 times for each participant, with each of the 10 folds serving as the held-out fold once, which implies each observation is predicted once out of sample.

Analysis

Datasets. Initially, we included in our analysis all 19 datasets assembled by He et al. (2022). However, upon further examination, we identified a discrepancy between the sequence of the problem IDs in the data used by He et al. and the problem IDs reported in three of the raw datasets, all from one experiment (Pachur, Schulte-Mecklenbeck, & Murphy, 2018, Table A3). Unfortunately, this meant there was no correspondence between choice tasks and choice rates in these three datasets, and models' performance was distorted. Additionally, in a fourth dataset (from Stewart et al., 2015) participants faced a substantial number of tasks twice, which implied those tasks appeared in both the training and test samples, and to data leakage. We thus excluded these four datasets from our analysis, leaving 15 datasets that include a total of 1565 choice tasks.[1]

The 15 datasets included in the current paper can be broadly divided into two main types of designs. Eight of the datasets originate from experiments that have manually altered the gamble design by systematically altering the distribution of gambles' payoffs and probabilities (e.g., Stewart et al., 2015). The other seven datasets come from experiments with a more representative environment, using only or mostly randomly generated problems (e.g., Rieskamp, 2008). Overall, there were 504 choice tasks in datasets with randomly generated tasks, and 1061 tasks in datasets with systematically generated tasks.

Prediction Error. We focus on prediction of the aggregate choice rates in each problem. Specifically, we averaged the predictions for all decision makers in a given problem t. Through this process, we obtained a single prediction of model m for each problem, denoted as $\hat{y}_{m,t}$. We then compute each dataset d's Mean Squared Error (MSE): $MSE_{m,d} = \frac{1}{N_d}\Sigma_{i=1}^{N_d}(\hat{y}_{m,t} - y_i)^2$ where y_i represents the observed choice and N_d is the number of problems in d. Finally, in our main analysis, we compare models based on their average MSE in a dataset (i.e., giving each dataset equal weight regardless of the number of problems it contains).

[1] Although this means that the comparison of the 58 original models as reported by He et al. (2022) is also partly flawed, the authors fortunately provided a detailed replication package that allows recalculating all scores without the four flawed datasets.

Results

Gains and Mixed Gambles

To assess the effectiveness of BEAST, we first repeated the primary analysis conducted by He et al. (2022) by comparing the predictive performance of the models on datasets containing only choice tasks in the gain domain separately from datasets containing tasks from the mixed gains/loss domain (Figure 1). Our results indicate that the original BEAST model (in blue) performed exceptionally well in the mixed gambles domain (Figure 1a), placing second, behind only one of the 58 models evaluated by He et al (2022). This result demonstrates that the exclusion of the model due to its difficulty to estimate led to the loss of a leading model in one of the two domains investigated in the original study.

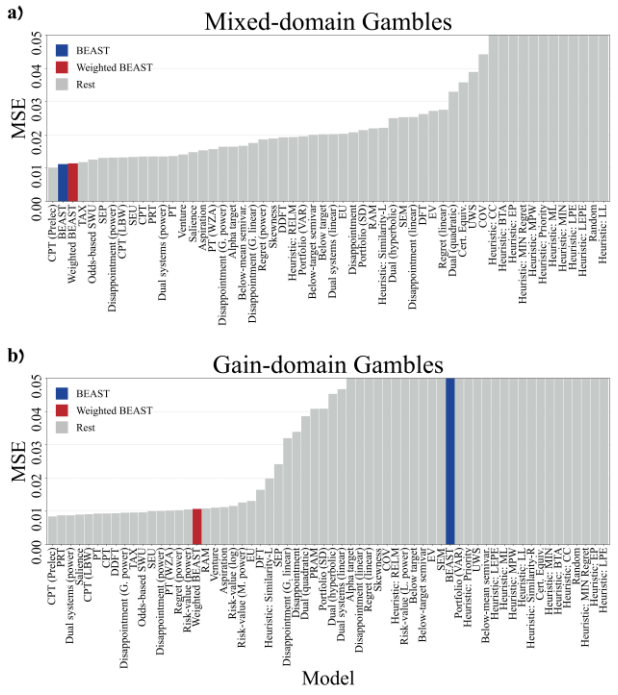

Figure 1: Average mean squared error of models in datasets with gambles in the Mixed-domain (a) and Gain-domain (b). Y-axis is truncated at 0.05 for convenience.

Despite the excellent predictive performance in the mixed domain, our analysis revealed that the original BEAST model struggled in the gain domain, showing quite poor predictive performance (Figure 1b). Analysis of the distribution of fitted parameters (see SM) suggests that in the gain domain, best fit of the original model reflects maximal attempt to account for deviations from maximization, under the constraint that the difference between EVs receives considerable weight.

To address this issue, we developed W-BEAST that relaxes this extreme constraint (as detailed in the method section). Our findings demonstrate that the modified model (in red) significantly improved the predictive performance of BEAST in the gain domain $[t_{(1,703)} = -23.39,\ p < 0.001]$, without compromising its performance in the mixed domain $[t_{(422.5)} = -0.055,\ p = n.s]$.[2]

Random and Systematic Datasets

We further analyzed our results by examining the performance of BEAST on individual datasets and found that its poor performance in the gain domain was largely driven by datasets by Stewart and colleagues (Stewart et al., 2015; Stewart, Hermens, & Matthews, 2016). The sets of choice tasks in these studies were specifically designed to elicit a context effect and alter decision-making behavior. This is in contrast to other datasets that included tasks that were randomly sampled from a large space of problems (e.g., Rieskamp, 2008). To better understand the effect of this task design issue on the performance of the model, we divided the datasets into two categories: datasets that mainly involved systematically crafted tasks and those that mainly involved randomly generated tasks.

Comparison of the models' predictions in the two types of datasets revealed that the original BEAST performed poorly on the systematic-tasks datasets (Figure 2b), but much better on random-tasks datasets (Figure 2a). Note that the latter now encompasses datasets from both the mixed-domain and the gain-domain. W-BEAST, as can be seen, has shown substantial improvements both in random-task datasets $[t_{(596.2)} = -3.43, p < 0.001]$ and in systematic-task datasets $[t_{(1,485)} = -23.47, p < 0.001]$. Further analysis by dataset (see SM) reveals that the revised model emerged as the winner on two random-task datasets (Erev et al., 2017; Rieskamp, 2008), one from the gain domain and the other from the mixed domain.

Sampling Tools

To shed more light on the components that facilitated the great improvement of W-BEAST in the systematic tasks, we analyzed its fitted weights of sampling tools.

One of the main constraints of BEAST is the assumption that each biased sampling tool (i.e., Uniform, Contingent Pessimism, and Sign) has equal probability to be selected. We examined the distribution of the fitted weights of the biased tools for all participants across all the folds, for the systematic-tasks and random-tasks separately. The variance in the weights was higher in the systematic-tasks datasets $[Var = .061]$ compared to the random-tasks datasets $[Var = .041]$, and the difference is significant using Levene's test $[F_{(1,\ 19,738)} = 411.7,\ p < 0.001]$. These results suggest that indeed constraining BEAST to give equal weights to each of the biased sampling tools hurt its performance more in the systematic tasks.

[2] To increase statistical power, these and the following statistical tests use problem as unit of analysis (comparing performance in predicting each task). Note that our main analysis (shown in the figures) compares the average performance in a dataset, but we repeated this analysis using problem as unit of analysis and got practically identical results.

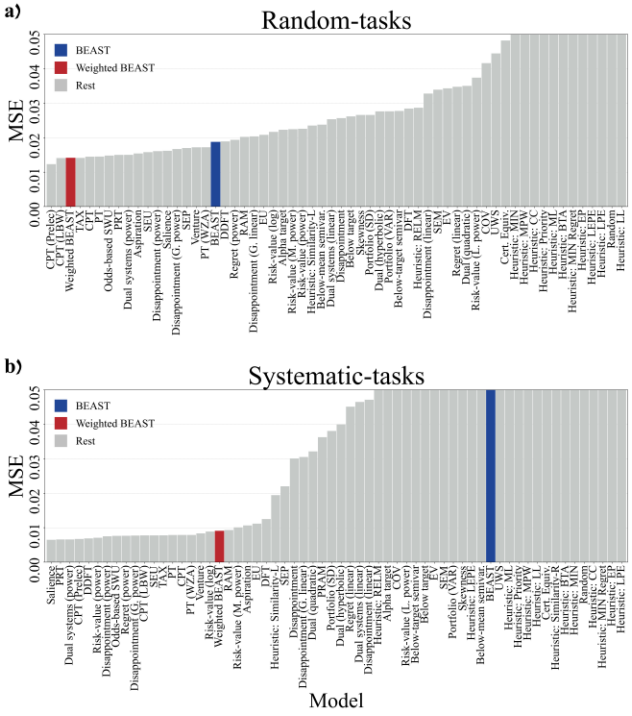

Figure 2: Average mean squared error of models in datasets with Randomly- (a) and Systematically- (b) generated tasks. Y-axis is truncated at 0.05 for convenience.

Additionally, the revised model's improved performance in the systematic-tasks datasets may be attributed to the over-sensitivity of BEAST to the difference between the gambles' EVs. Because relying on many mental draws through the unbiased tool is similar to giving large weight to the EV, we compared the weight of this sampling tool between the random- and the systematic-tasks datasets. Indeed, the weight assigned to the unbiased tool in the random-tasks $[M = .507, SD = .267]$ was significantly greater than in the systematic-tasks $[M = .328,\ SD = .281;\ T_{(6,578)} = 24.47,\ p < 0.001,\ Cohen's\ d = 0.654]$. These results help explain why the predictions made by the original BEAST are far more accurate in the random-tasks datasets compared to the systematic-tasks datasets, and highlight how relaxing the strong constraints imposed in the original implementation of BEAST allow it to significantly improve predictions when tasks are not chosen in a representative manner.

Discussion

This study demonstrates the potential of the BEAST model in decision-making research. Our analysis revealed that the original BEAST model achieved very strong performance in the domain of mixed gambles, placing among the top two models in the study conducted by He et al. (2022). However, the model struggled in the gain domain, showing poor predictive performance. To address this, we developed W-BEAST, a modified flexible version of BEAST, which resulted in an immense improvement in its performance in the gain domain while maintaining its performance in the mixed domain. Furthermore, W-BEAST demonstrated excellent results on a larger batch of datasets with a strong orientation for randomly generated problems, consisting of both mixed domain and gain domain datasets.

The results of our analysis align with the original intent for which the BEAST model was developed. The model was specifically designed to capture a broad spectrum of phenomena in human behavior and predict human decision-making in wide sets of environments (Erev et al., 2017). In this study, we evaluated the model's performance separately on datasets with randomly generated choice tasks that are arguably more representative of general decision making environments. Our findings demonstrate that the model exhibited excellent performance on these randomly generated tasks, thereby affirming its ability to accurately capture the natural variability of decision-making problems and predict human behavior in a diverse range of scenarios. This is further supported by the fact that among all the models tested, W-BEAST performed the best on the only two datasets that involved only randomly generated tasks.

As the original model was developed to capture behavior in wide classes of decision making tasks (including decisions under ambiguity and, primarily, following feedback), its developers chose to implement it with multiple constraints that limit its flexibility but reduce the number of free parameters and risk of overfitting. These additional constraints have prevented BEAST from capturing behavior in several of the datasets that include systematically crafted choice tasks. We conjecture that the reason for this is that the systematic gamble design has led to large context effects. For example, in one of the tasks from Stewart, Hermens and Matthews (2016), participants were asked to choose between a sure gain of 100 and a gamble that provided 500 with probability 0.9. The choice rate of the gamble was only 50%, despite the very large difference in expected values. This example reflects a general pattern in that dataset. Behavior reflecting this level of risk aversion is probably a result of an idiosyncratic context effect in that particular experiment and is very unlikely to be general. BEAST that gives significant weight to the difference between expected values cannot capture this type of behavior. More generally, it is not surprising that the inflexible model designed to capture behavior in general decision making contexts struggles when applied to datasets that include large idiosyncratic effects. To tackle this issue, we introduced a modified version of BEAST that preserves its original mechanisms and choice strategies. Our results show that W-BEAST demonstrates a substantial improvement in the predictive performance in datasets with systematically chosen choice tasks. Interestingly, the model also improves when predicting randomly generated tasks. These results suggest that an amended version of BEAST, like the one we present here, which allows it more flexibility, can increase the model's robustness to different types of data and its relevance also in specific contexts.

The results of this study have important implications for the field of decision-making research. It is crucial to consider

the unintended consequences of disregarding non-analytic models, such as BEAST, in decision-making research. The case we present here serves as an evident example of how such models, despite the difficulty of their estimation, can display strong results in decisions under risk. By expanding the scope of models examined in decision-making research, we can gain a more nuanced and accurate comprehension of the complexities of human behavior in decision-making contexts.

In the case of BEAST, the exclusion of the model may have led to an undervaluation of the underlying psychological mechanisms that drive the decision making process of the model. In their paper, He et al. (2022) found that payoff and probability transformations have much larger contributions to predictive performance than other mechanisms. The conclusion that payoff and probability transformations are key mechanisms for predictive performance is largely derived from the fact that the majority of top-performing models in the mixed gambles domain utilize such mechanisms. However, BEAST operates primarily through different mechanisms and specifically sampling and regret. On top of that, He et al. demonstrated that the best crowd models, which outperform the 58 individual models by weighting them, assigned relatively large weights to specific models that, while reasonably successful, employ unique mechanisms not shared by the top performing models. This finding indicates that incorporating a model like BEAST that relies on very different assumptions than those common in mainstream models like Cumulative Prospect Theory (CPT; Tversky & Kahneman, 1992) and that predicts out of sample well, could greatly enhance prediction accuracy of crowd models.

One of the goals in comparing predictive performance of models is to enhance our ability to predict the behavior of people in the real world. For example, highly accurate predictive models of human choice can be used to simulate humans when training artificial agents that would later be deployed in the wild (e.g., Moisan & Gonzalez, 2017; Raifer et al., 2022). However, in the study conducted by He et al. (2022), models were trained and tested on the same sample of participants (although predicting behavior in choice tasks models were not trained on). This raises the question of the generalizability of the results to new samples or the population at large. For example, the best performing model is a specific version of CPT (Prelec, 1998), which recently has been found to be a very unrestrictive and flexible model of choice under risk (Fudenberg, Gao, & Liang, 2020). Specifically, Fudenberg et al. find that CPT gives very good out of sample predictions for real data, but is sufficiently flexible so that it "would have performed well out-of-sample given sufficient data from almost any underlying data-generating process that respects first-order stochastic dominance" (Fudenberg et al., 2020, p. 21). This suggests that the individually fitted CPT is likely to capture any patterns a participant displays in an experiment, regardless of how well the underlying choice process is reflected by the main assumptions of the theory. When the goal of the predictive model is to mimic the future expected behavior of the individual to which the model is fitted, using such flexible model can be useful. But, when predicting for new samples, this type of flexibility may be problematic. Flexible models may overfit to the individuals in-sample, and the merits of low-restrictive models on the individual level can become drawbacks when predicting for the population or other samples. BEAST, on the other hand, was originally designed to predict choice behavior at aggregate levels and its assumptions are notably more restrictive. In light of this, in future work, it can be beneficial to assess the predictive capabilities and generalizability of the models presented in He et al. (2022), with the addition of BEAST and W-BEAST, by testing them on similar datasets obtained from new samples of participants. Such comparison may then also be a better test of the underlying assumptions and theories behind the different models.

Although too much flexibility can sometimes be a concern, our results highlight that too little flexibility is also not ideal. The inflexible BEAST performed poorly when applied to systematic-task datasets. W-BEAST that decreased this level of inflexibility of the model has made large improvements in predicting the possibly idiosyncratic behavior reflected in those datasets. An examination of the model's fitted parameters revealed that the original BEAST's assumptions of equal probability to use each biased sampling tool and high sensitivity to the difference between expected values were too restrictive. We found higher variance in the weights of the biased sampling tools for systematic tasks, pointing to a stronger dependence on a single biased sampling tool. Conversely, in random tasks, all three biased tools were estimated to be used with roughly equal probability, akin to the original assumption of the BEAST. We further found reduced reliance on the unbiased tool (suggesting lower sensitivity to the EVs) for systematic tasks compared to random tasks. These results imply that the use of biased sampling tools is particularly important when producing predictions for systematically crafted tasks. In future work, it would be beneficial to consider incorporating other biased sampling tools and evaluating their usefulness in contextualized settings.

In sum, this study highlights the predictive power of BEAST for choice under risk. Our analysis revealed that W-BEAST performed exceptionally well on a large batch of randomly generated tasks. These results accentuate the model's robustness and versatility, making it relevant in a wide range of scenarios. More importantly, it underscores the importance of considering a more inclusive approach when evaluating quantitative models in decision-making research, even if they may be considered more complex to estimate. Doing so can facilitate a deeper understanding of the underlying mechanisms that drive human behavior in decision-making contexts, which can lead to the development of more accurate models, improved decision-making strategies, and greater understanding of the psychological processes involved in decision-making.

Acknowledgments

Ori Plonsky acknowledges support from the Israel Science Foundation (grant no. 2390/22).

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'ecole ´ americaine. ´ *Econometrica: Journal of the Econometric Society* 21(4), 503–546. https://doi.org/10.2307/1907921

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, 113(2), 409–432. https://doi.org/10.1037/0033-295X.113.2.409

Bugbee, E. H, & Gonzalez, C. (2022). Making Predictions Without Data: How an Instance-Based Learning Model Predicts Sequential Decisions in the Balloon Analog Risk Task. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44.

Cousineau, D., & Allen, T. A. (2015). Likelihood and its use in parameter estimation and model comparison. *Mesure et Evaluation en Éducation*, 37(3), 63–98. https://doi.org/10.7202/1036328ar

Edwards, W. (1954). The theory of decision making. *Psychological bulletin*, 51(4), 380-417. https://doi.org/10.1037/h0053870

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75(4), 643-669. https://doi.org/10.2307/1884324

Erev, I., Ert, E., & Roth, A. E. (2010). A choice prediction competition for market entry games: An introduction. *Games*, 1(2), 117-136. https://doi.org/10.3390/g1020117

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, 124, 369-409. https://doi.org/10.1037/rev0000062

Ert, E., & Erev, I. (2013). On the descriptive value of loss aversion in decisions under risk: Six clarifications. *Judgment and Decision Making*, 8(3), 214–235.

Fudenberg, D., Gao, W., & Liang, A. (2020). How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories. *arXiv preprint*. https://doi.org/10.48550/arXiv.2007.09213

He, L., Analytis, P. P., & Bhatia, S. (2022). The wisdom of model crowds. *Management Science*, 68(5), 3635-3659. https://doi.org/10.1017/S1930297500005945

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica* 47(2), 263–292.

Moisan, F., & Gonzalez, C. (2017). Security under uncertainty: adaptive attackers are more challenging to human defenders than random attackers. *Frontiers in psychology*, 8. https://doi.org/10.3389/fpsyg.2017.00982

Pachur, T., Schulte-Mecklenbeck, M., Murphy, R. O., & Hertwig, R. (2018). Prospect theory reflects selective allocation of attention. *Journal of experimental psychology: general*, 147(2), 147-169. https://doi.org/10.1037/xge0000406

Payne, J. W. (2005). It is whether you win or lose: The importance of the overall probabilities of winning or losing in risky choice. *Journal of Risk and Uncertainty*, 30, 5–19. https://doi.org/10.1007/s11166-005-5831-x

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. Psychological review, 122(4), 621-647. https://doi.org/10.1037/a0039413

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., ... & Erev, I. (2019). Predicting human decisions with behavioral theories and machine learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.1904.06866

Plonsky, O., & Erev, I. (2021). Prediction oriented behavioral research and its relationship to classical decision research. *PsyArXiv*. https://doi.org/10.31234/osf.io/7uha4

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3), 497-527. https://doi.org/10.2307/2998573

Raifer, M., Rotman, G., Apel, R., Tennenholtz, M., & Reichart, R. (2022). Designing an automatic agent for repeated language–based persuasion games. *Transactions of the Association for Computational Linguistics*, 10, 307-324. https://doi.org/10.1162/tacl_a_00462

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6), 1446-1465. https://doi.org/10.1037/a0013646

Stewart, N., Reimers, S., & Harris, A. J. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, 61(3), 687-705. https://doi.org/10.1287/mnsc.2013.1853

Stewart, N., Hermens, F., & Matthews, W. J. (2016). Eye movements in risky choice. *Journal of behavioral decision making*, 29(2-3), 116-136. https://doi.org/10.1002/bdm.1854

Thorngate W. (1980). Efficient decision heuristics. *Behavioral Science*. 25(3), 219–225. https://doi.org/10.1002/bs.3830250306

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5, 297-323. https://doi.org/10.1007/BF00122574