Judgment and Decision Making (2024), **vol. 19**:e35 1–16
doi:10.1017/jdm.2024.34

EMPIRICAL ARTICLE

Beyond analytic bounds: Re-evaluating predictive power in risky decision models

Or David Agassi and Ori Plonsky

The Faculty of Data and Decision Sciences, Technion – Israel Institute of Technology, Haifa, Israel

Corresponding author: Ori Plonsky; Email: plonsky@technion.ac.il

Received: 19 November 2023; **Revised:** 9 September 2024; **Accepted:** 9 September 2024

Keywords: decisions under risk; choice prediction; model comparisons; computational modeling

Abstract

Research in behavioral decision-making has produced many models of decision under risk. To improve our understanding of choice under risk, it is essential to perform rigorous model comparisons over large sets of decision settings to find which models are most useful. Recently, such large-scale comparisons have produced conflicting conclusions: A variant of cumulative prospect theory (CPT) was the best model in a study by He, Analytis, and Bhatia (2022), whereas variants of the model BEAST were the best in two choice prediction competitions. This study delves into these contradictions to identify and explore the underlying reasons. We replicate and extend the analysis by He et al., this time incorporating BEAST, which was previously excluded because it cannot be analytically estimated. Our results show that while CPT excels in systematically hand-crafted tasks, BEAST—originally designed for broader decision-making contexts—matches or even surpasses CPT's performance when choice tasks are randomly selected, and predictions are made for new, unknown decision makers. This success of BEAST, very different from classical decision models—as it does not assume, for example, subjective transformations of outcomes and probabilities—puts into question previous conclusions concerning the underlying psychological mechanisms of choice under risk. Our results challenge the field to expand beyond established evaluating techniques and highlight the importance of an inclusive approach toward nonanalytic models, like BEAST, to achieve more objective insights into decision-making behavior.

Research in judgment and decision-making often observes clear deviations from the predictions of normative models of choice under risk and uncertainty like expected utility theory. This has led to the development of many so-called descriptive models, meant to describe how people actually make choices in risky and uncertain situations (e.g., Kahneman & Tversky, 1979; Tversky & Kahneman, 1992; Savage, 1954; Bell, 1982). As these models proliferate, determining their relevance in various decision contexts emerges as a pivotal challenge, underscoring the need for systematic model evaluations and comparisons.

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 answered by large samples of participants. 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 numbers of parameters, and increases the chances that diverse models will be developed (Plonsky & Erev, 2021a).

In a recent impressive study, He, Analytis, and Bhatia (2022), hereafter HAB, performed a large-scale comparison of dozens of models of decision-making under risk. They grouped 19 datasets from multiple published studies with more than 1800 choice tasks, each a one-shot decision between two fully described gambles with up to two possible outcomes. They analyzed both datasets with tasks involving only potential gains (hereafter the gain domain) and datasets with tasks involving both gains and losses (hereafter 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 (e.g., Allais, 1953; Erev et al., 2017, Ert & Erev, 2013; Kahneman & Tversky, 1979; Stewart et al., 2015). HAB compared 58 published models of risky choice to examine which of these offers the best predictions of individual decision makers' choices for this large data. The results revealed that a variant of cumulative prospect theory (CPT; Prelec, 1998) was the best predictive model in both the gain domain and the mixed domain.[1]

However, in two other recent large-scale model comparison studies, Choice Prediction Competitions (CPC) 2015 (Erev et al., 2017) and 2018 (Plonsky et al., 2024), CPT did not fare as well. In these competitions, any model, including CPT, could be independently submitted and evaluated for its prediction accuracy. The results of the competition showed that the model BEAST (Best Estimate and Sampling Tools; Erev et al., 2017), which is very different from mainstream models like CPT, emerged as the one with the most accurate predictions.[2] It is thus of interest to investigate the reasons behind these differing results.

1.1. Overlooking nonanalytic models

Aside from the different winning models, several differences exist between the study conducted by HAB and the two CPCs. The most significant difference is HAB's decision to exclude a certain category of models from their large-scale analysis. Specifically, they chose to exclude models that could not be fitted easily using analytical likelihood functions (hereafter nonanalytic models), like those that require running simulations to make predictions. This choice, incidentally, implied the exclusion of the model BEAST from the analysis. The decision to exclude nonanalytic models reflects a general practice in the field that tends to focus on models that are amenable to estimation and those with easily identifiable parameters. The focus on such models diminishes modeling 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, 2021a). However, there is no apriori reason to assume that a theoretical "ideal model" of decision-making must necessarily fall within the space of models that are easily estimable. Ignoring nonanalytic models may 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 providing highly useful predictions of behavior (Erev et al., 2010; Erev et al., 2017; Plonsky et al., 2024).

Furthermore, nonanalytic models are often implemented in ways that are not amenable to easy estimation because they assume psychological processes that are hard (or impossible) to implement analytically. Disregarding the potential of nonanalytic models thus may lead to wrong conclusions about the underlying psychological processes that are important for choice prediction. For example, (He et al., 2022) concluded that subjective nonlinear payoff and probability transformations are

[1] When the prediction error was defined as the mean squared error, the measure we focus on in this study.

[2] Three variants of CPT were submitted to CPC15 (Erev et al., 2017), all of which predicted significantly worse than BEAST, with the best of the three achieving 187% the error of BEAST. No CPT variants were submitted to CPC18 (Plonsky et al., 2024), but the organizers fitted and evaluated CPT (on the subset of the data that includes only choice under risk without feedback) and found that its error is roughly twice that of BEAST.

https://doi.org/10.1017/jdm.2024.34 Published online by Cambridge University Press

essential mechanisms for choice prediction accuracy. However, BEAST does not assume either of these mechanisms. Rather, it uses simulations to derive the predictions of a process of mental sampling of potential outcomes which is sensitive to anticipated regret. Hence, evaluating BEAST can help shed light on the usefulness of assuming very different psychological processes than mainstream models assume.

In this paper, we seek to reconcile the inconsistent results between the two CPCs, where BEAST emerged as the superior predictive model, and HAB's comparison, where CPT was identified as the leading model. To do so, we apply identical methods to those used by (He et al., 2022) and examine the predictive power of BEAST, which was excluded from their analysis, on their data. Notably, HAB's data includes only one-shot binary decisions under risk with up to two outcomes, whereas BEAST was originally developed to capture choice in a much wider class of tasks (including decisions under ambiguity and decisions under risk with repeated feedback). The wide applicability of BEAST has led its developers, concerned with overfitting issues, to introduce several arbitrary implementation assumptions that restrict the model in ways that are not necessarily implied by the underlying theory, but save free parameters. In the simple domain of choice under risk with up to two outcomes, BEAST requires less free parameters. This allowed us to also develop and test a version of BEAST that relaxes some of the original restrictive implementation assumptions. This version, which we call AdaBEAST, maintains the underlying psychological rationale of BEAST but allows for increased adaptability to different study contexts. Our study thus investigates how two nonanalytic models, BEAST and AdaBEAST, fare in comparison to dozens of analytic choice models in one of the most basic decision tasks, and explores what can be learned from this comparison.

2. The structure of BEAST and its distinction from classical models

Before presenting the analysis, it is useful to clarify the main theoretical underpinnings of BEAST, its main mechanistic structure, and the main differences from mainstream choice under risk models like CPT. (Implementation details are left for the Methods section and the Supplementary Material.)

BEAST belongs to a class of models that rely on the conjecture that judgment and decision-making reflect cognitive strategies or tools that have been effective in past experiences perceived as similar to the current situation (e.g., Plonsky et al., 2015). A key assumption, which builds on Skinner's notion of "contingencies of reinforcement" (Skinner, 1953), is that people act as "intuitive classifiers" (Erev & Marx, 2023). They use environmental cues to intuitively determine the class that the current situation belongs to, and then rely on strategies that previously worked well in this class of situations. This subjective imperfect classification process is often effective, but can also lead to behavioral biases: When people misclassify a situation to a class that only appears similar, they rely on strategies that can be counter-productive in the current context (Erev et al., 2017).

This cognitive process that relies on potentially intricate subjective similarity relationships can be highly complex, and extant models (including BEAST) often do not represent it explicitly. Rather, models that rely on this conjecture aim to approximate the main implications of the complex process. Past research shows that this can be done by surprisingly simple models (Erev & Roth, 2014; Erev et al., 2023) that assume people behave as if they take small mental samples of potential outcomes of the possible actions and tend to choose the action with the best average outcome in the sample (cf., e.g., Juslin et al., 2007; Vul et al., 2014; Zhu et al., 2020). Many of the previous models in this class were developed to capture repeated choice behavior. Hence, it was natural to assume that the sampled outcomes are drawn from the payoffs observed in previous experiences with the same task. BEAST, in contrast, aims to (also) capture initial (pre-feedback) choice behavior. The sampled outcomes are thus assumed to reflect the results of cognitive strategies that worked well in situations outside the current experimental context. These strategies are implemented as potentially biased "sampling tools."

To evaluate a choice option, BEAST agents mentally sample possible payoffs from either the objectively described payoff distribution—using the so-called *unbiased* sampling tool—and/or from transformed (i.e., biased) versions of that distribution. The use of the three biased sampling tools in

https://doi.org/10.1017/jdm.2024.34 Published online by Cambridge University Press

BEAST— *contingent pessimism*, *uniform*, and *sign* —aims to capture reliance on past experiences where the information provided was not accurate and objective but biased or irrelevant. Specifically, the contingent pessimism tool implies sampling (with certainty) the worst payoff described. Its use may reflect a subjective perception of a task as similar to situations where an adversary could influence the agent's realized payoff. A tendency to perceive tasks pessimistically can help explain behavioral phenomena like loss aversion (Samuelson, 1963), the Allais paradox (Allais, 1953), and the St. Petersburg paradox (Bernoulli, 1954). The uniform sampling tool implies an equiprobable sampling of one of the described payoffs. Its use may reflect subjectively classifying a task as similar to situations where ignoring probability information (which, unlike payoff information, is often unverifiable even after the fact) was beneficial. A tendency to rely only on payoff information in this manner can help explain phenomena like overweighting of rare events (Friedman & Savage, 1948) and the Allais paradox. Finally, the sign sampling tool involves unbiased sampling of payoffs after they have been transformed using the sign function. Its use may reflect a subjective perception of a task as similar to contexts where the main goal was to avoid losses (e.g., the initial rounds of survival tournaments where the worst performers are eliminated). A tendency to focus on the payoff sign can help explain phenomena like the reflection effect (Markowitz, 1952).[3]

The mental sampling mechanism in BEAST implies a very different process than mainstream models of choice under risk, like CPT, and has very different implications. Mainstream models, like CPT, capture deviations from expected value (EV) maximization like those mentioned above by assuming they reflect subjective transformations of each possible payoff. These are then weighted together by subjective transformations of their respective probabilities, thus creating a subjective utility of a given choice option. Most commonly, these transformations are nonlinear but monotonic, such that higher and more probable payoffs necessarily contribute more to the subjective utility. In contrast, in BEAST, due to the inherent stochasticity in the mental sampling process, it is plausible that some payoffs may be overweighted, underweighted, or even neglected entirely. Furthermore, depending on the sampling tool that is applied, higher or more probable payoffs might contribute similarly or less to the decision than other payoffs.

Moreover, in most mainstream choice models, including CPT, the evaluation of each choice option, and its subjective utility, is formed independently of the alternative options. In contrast, the use of sampling tools in BEAST is a function of the choice task, not of particular choice options. For example, if a task is perceived as (also) similar to adversarial situations, then the contingent pessimism tool will be used on all options. Furthermore, the outcomes are sampled from the choice options in a correlated manner and are often directly influenced by the properties of the alternative choice options. This correlated sampling process implies that BEAST in essence includes a mechanism of anticipated regret from choosing one option over the other: The attractiveness of one choice option is a function of the expected attractiveness of the other choice option. This type of context dependence was recently suggested to be crucial for useful choice prediction (Peterson et al., 2021; Plonsky & Erev, 2021b), but is absent in many classical models.

3. Method

The current study includes four sets of analyses that aim to replicate and extend the analyses conducted by (He et al., 2022). First, we replicate HAB's use of an ensemble of choice datasets to examine how well dozens of models of decisions under risk predict the decisions made by familiar individuals in

[3] In addition to the assumption of mental sampling, BEAST assumes choice is sensitive to the difference between the options' expected values. This assumption was introduced to the model based on empirical evidence, but it is not easily justifiable under a framework of intuitive subjective classification. In the modified version of BEAST that we propose in this article, we remove this assumption.

https://doi.org/10.1017/jdm.2024.34 Published online by Cambridge University Press

new (out-of-sample) choice tasks. The term “familiar” is used here to highlight that the models were evaluated based on choice data from the same individual decision makers on whom the models were also estimated (but using different choice tasks). Importantly, we now add to this examination two nonanalytic models, BEAST and its modified version, AdaBEAST. Second, as HAB, we analyze the psychological mechanisms that are embedded in the most successful models, aiming to understand better the underlying choice processes. Third, we investigate the predictive power of the different models in each of the different datasets that are part of the ensemble used in the main analysis. The aim here is to identify whether specific features of the dataset are associated with greater success of specific models. Finally, we extend the analysis by evaluating the predictive power of the models on “unknown” (rather than “familiar”) out-of-sample individuals. Specifically, we assess the models’ ability to predict the choices—in new choice tasks as before—of decision makers whose data was *not* included in the model estimation phase. Under this approach, neither the participants nor the tasks that models are required to predict are seen by the models during the model fit, testing the models’ ability to generalize to new decision makers and new decision contexts.

3.1. Models

Details for all 58 “analytic” models in the model comparison and their estimation procedures can be retrieved from (He et al., 2022). As mentioned, we add to the comparison two nonanalytic models: BEAST and AdaBEAST.

3.1.1. BEAST

The following describes the main implementation assumptions of BEAST that are relevant to the current investigation. More complete details can be found in the supplementary material (SM) as well as in Erev et al., 2017.

In binary decision under risk problems, BEAST 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 their 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, it is assumed $e = 0$.

ST is the average of κ_i outcomes that are each mentally sampled using one of four possible sampling tools. In each of the κ_i independent sampling instances, two outcomes, one from each option, are sampled using the same sampling tool, and under the assumption that the payoff distributions from which sampling takes place are positively correlated (a “luck level” procedure – see SM for details). This implies high sensitivity to the anticipated regret, defined as the difference between the outcome sampled in one option and the one sampled from its alternative.

Sampling tool *unbiased* implies unbiased draw from the options’ described distributions. The remaining three sampling tools imply biased sampling. The sampling tool *uniform* ignores the described probabilities and samples as if all potential outcomes are equally likely (Thorngate, 1980). Sampling tool *contingent pessimism* yields the worst possible outcome (Edwards, 1954) under some lexicographic conditions (Brandstätter, Gigerenzer & Hertwig, 2006) that depend on some value γ_i and on the ratio between the worst outcomes of the two options (see SM for details). The sampling tool *sign* is similar to the unbiased tool, but samples only the sign of the outcome, ignoring magnitudes (Payne, 2005). BEAST assumes that the probability to use each of the three biased sampling tools is the same and that the probability of using the unbiased tool is $1 - \frac{\beta_i}{\beta_i+1}$.

Finally, an individual decision maker i’s parameters are assumed to be drawn from uniform distributions as follows: $\sigma_i \sim U\left[0, \sigma\right]$, $\kappa_{\mathrm{i}} \sim \mathrm{U}\left(1, 2, \ldots, \kappa\right)$, $\gamma_i \sim U\left[0, \gamma\right]$, $\beta_i \sim U\left[0, \beta\right]$ with $\sigma, \kappa, \gamma, \beta$ free parameters to be estimated.

https://doi.org/10.1017/jdm.2024.34 Published online by Cambridge University Press

3.1.2. AdaBEAST

The original 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 and avoid overfitting, its developers made several rather arbitrary implementation assumptions that heavily restrict the model but save free parameters. Since here we focus on one-shot decisions under risk with up to two outcomes without feedback (the setting investigated by (He et al., 2022)), we developed a modified, more adaptable version of the model, AdaBEAST, which relaxes some arbitrary restrictions yet preserves the main logic and mechanisms underlying BEAST. The following presents details on the changes from BEAST.

One highly restrictive (and theoretically arbitrary) assumption embedded in BEAST is that each of the biased sampling tools is used with the same probability. Clearly, however, different choice contexts may lead decision makers to perceive the choice tasks as more similar to some situations than to others, which implies different likelihoods for using different sampling tools. To capture possible idiosyncratic contextual effects of different datasets, AdaBEAST relaxes this restrictive assumption. In AdaBEAST, the probability of using each of the biased sampling tools is a free parameter. Specifically, W_{uf}, W_s and W_{cp} represent the probability of using the *uniform, sign*, and *contingent pessimism* tools respectively. The probability to use the *unbiased* sampling tool is then simply $W_{ub} = 1 - \left(W_{uf} + W_s + W_{cp}\right)$.

Another restrictive assumption originally implemented in BEAST is that the difference in averages of the mental samples taken $(ST_A - ST_B)$ is equally weighted with the difference between the options' EVs $(EV_A - EV_B)$. Sensitivity to the difference between EVs was originally introduced to BEAST based on empirical evidence that choice is highly correlated with EV maximization. Yet, it is not clear how it fits within a framework of intuitive subjective classification that BEAST derives from. Further, even if good models of choice under risk should reflect sensitivity to EV difference, the choice to weight it equally to the output of the mental sampling process is highly restricting and rather arbitrary (but saves a free parameter). To preserve the model's possibility to account for high rates of EV maximization, remain within the general framework of mental sampling as a reflection of intuitive subjective classification, and increase the model's adaptability, we chose to make two changes when developing AdaBEAST.

First, the size of the mental sample taken from each option, κ_{i}, is assumed to be drawn from a Geometric distribution with parameter p (free parameter). That is, $\Pr(\kappa_i = k) = (1-p)^{k-1}p$. This change is based on the observation that most decision makers behave as if they rely on small samples (e.g., Plonsky et al., 2015), but allows for some to behave as if they rely on large ones (Erev et al., 2023). Second, we completely removed the fixed dependence on EV difference. That is AdaBEAST implies option A is preferred over option B if $[ST_A - ST_B] + e > 0$. Note that because the weight of the unbiased sampling tool in AdaBEAST (W_{ub}) can now range from 0 to 1 and κ_i can be large (which is more likely when p is small), it is possible for AdaBEAST to rely on an approximation of the EV: a large unbiased sample from the outcome distribution. Hence, AdaBEAST can still capture the behavior that appears as sensitivity to the differences between EVs without having to assume such sensitivity explicitly.

The differences between BEAST and AdaBEAST are at the level of implementation assumptions, not at the level of the underlying mechanisms and logic. AdaBEAST preserves the idea of subjective intuitive classification explained above, as well as the idea that in choice under risk, the implications of this intuitive classification process can be summarized by the use of four sampling tools. Yet, AdaBEAST arguably implements the process in a more natural and cognitively plausible manner as it allows the environment to influence the likelihood for each classification and avoids explicit computations of the EVs. Python code for AdaBEAST can be found in the SM (see https://osf.io/ca6bn/).

3.2. Data

The original data analyzed by (He et al., 2022) included 19 distinct datasets. However, upon inspection, we found that four of these were not usable for this analysis. Three of the datasets, all from the same experiment (Pachur et al., 2018), had discrepancies between the raw data used by HAB and the actual choice rates as reported in the original article by Pachur et al. (2018; Table A3). The source of these discrepancies lies in inconsistencies between task IDs in the raw data and the original task IDs, which, unfortunately, led to distorted computed model performances in HAB's analysis. In a fourth dataset (from Stewart et al., 2015), participants faced a substantial proportion of the choice tasks twice in the same session, which implied those tasks appeared in both the training and test samples, leading to data leakage. We thus excluded these four datasets from our analysis, leaving 15 datasets that include a total of 1565 choice tasks published in: Erev et al. (2017), Fiedler & Glöckner (2012), Pachur et al. (2017; 2018), Rieskamp (2008), and Stewart et al. (2015; 2016).[4]

Each dataset includes choice data from a different experimental context. Participants (sample sizes range from 30 to 208) in each context made multiple one-shot binary choices between lotteries with up to two possible outcomes each. The lotteries' payoff distributions were fully and accurately described and participants did not receive any feedback on their choices. Number of tasks per dataset (and participant) ranged from 50 to 150. Figure S1 in the SM shows an example of a choice task from one experiment.

3.3. Estimation and cross-validation

Fitting the models to the new data requires the estimation of the parameters $\sigma, \kappa, \gamma, \beta$ for BEAST and the parameters $p, \sigma, W_{uf}, W_s, W_{cp}$ for AdaBEAST. Because the models are simulation-based and do not have a differentiable likelihood function, we performed a grid search to find the best set of parameters. Specifically, in each dataset, we first generated the models' predictions for each choice task and each profile of parameters implied by the grid. We then estimated the models using a cross-validation technique similar to that used for the other 58 models in the original study by (He et al., 2022). Each participant's choice data was split into the same exact 10 folds of choice tasks as in HAB. In each cross-validation iteration, we chose the profile of parameters that best fits 9 of these folds (training data, representing 90% of the choice tasks in each dataset), based on the maximum likelihood criterion (Cousineau & Allen, 2015), and then elicited the fitted models' predictions for the held-out fold (test data, 10% of the choice tasks in each dataset). 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 the sample. The SM includes further details on the grid search fitting procedure.

3.4. Analysis

3.4.1. Prediction error

Similar to (He et al., 2022), we focus on the prediction of the choices of individual decision makers in each task. In the main analysis, for each individual i, each model m is fitted to the training choice data of that individual and generates a prediction $\widehat{y}_{i,m,t}$ for each out-of-sample task t. We then compute, for each individual, the mean squared error (MSE) of the individually fitted model across all tasks: $MSE_{m,i} = \frac{1}{N}\Sigma_{t=1}^{N}\left(\widehat{y}_{i,m,t} - y_{i,t}\right)^2$ where $y_{i,t}$ represents the observed choice of individual i in task t and N is the number of tasks the individual faced. Finally, we compare models based on their average MSE across all individuals (i.e., giving each individual equal weight regardless of the number of tasks he or she faced).

To extend the previous analysis, we performed an additional comparison aimed at assessing the prediction error models make for an unknown, out-of-sample individual i_{new}. Here, models are not trained on any of the choice data produced by this individual. To create a model m's prediction for

[4] Although the exclusion of the four unusable datasets means that the original comparison of the 58 models as reported by He et al. (2022) is also partly flawed, the authors fortunately provided a detailed replication package (see https://pubsonline.informs.org/doi/suppl/10.1287/mnsc.2021.4090) that allows recalculating all scores without the four flawed datasets. The results we report here concern performances on the 15 remaining datasets.

https://doi.org/10.1017/jdm.2024.34 Published online by Cambridge University Press

i_{new} in task t, we average the predictions the model makes for all other participants in the dataset who faced the same task, excluding the target individual i_{new}. That is: $\widehat{y}_{i_{new},m,t} = \frac{1}{n-1}\sum_{i\neq i_{new}} \widehat{y}_{i,m,t}$, with n the number of participants in the dataset.[5] The MSE for each out-of-sample individual is then calculated as before, and we report the average of these MSEs.

To check for statistical significance between the prediction errors of any two behavioral models, we implemented (using packages lme4, Bates et al., 2014, and lmerTest, Kuznetsova et al., 2017, in R) a linear mixed-effects statistical model with a fixed effect for the behavioral model and random intercepts for participants and for cross-validation fold of a dataset. We use the Satterthwaite approximation (Satterthwaite, 1946) to compute degrees of freedom.

3.4.2. Psychological mechanism classification

As part of the large-scale comparison of risky choice models, (He et al., 2022) classified the evaluated models as having or not each of nine different psychological mechanisms: payoff transformation, probability transformation, attention, sampling, regret, disappointment, ranking, threshold, and dispersion. We use HAB's classification for all models that they evaluated. BEAST and AdaBEAST we classify as involving both sampling and regret but none of the other mechanisms. The inclusion of sampling and regret follows from the description of the models provided above. Concerning the exclusion of other mechanisms, it may be argued that since the uniform sampling tool assumes outcomes are sampled uniformly, the models include a mechanism of probability transformation (e.g., to 0.5 in 2-outcome gambles). Indeed, this tool allows BEAST to capture the *behavior* that appears as if small probabilities are overweighted nonlinearly. Yet, the essence of a nonlinear probability transformation mechanism, as understood in almost every case, is the consistent nonlinear treatment of described probabilities. In contrast, in BEAST and AdaBEAST, the uniform sampling tool, which may not even be used in the decision process, never even considers the objective probabilities, let alone transforms them (indeed, it operates identically even when probabilities are unknown to the agents). Hence, in our analysis, we do not consider these models as involving a nonlinear probability transformation mechanism.

4. Results

4.1. Replicating He, Analytis, & Bhatia (2022)

To assess the effectiveness of BEAST, we first repeated the primary analysis in (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 in the mixed domain (Figure 1). Our results indicate that relative to all other behavioral models, the original BEAST model (thin arrow) demonstrated decent predictive performance in the mixed domain (Figure 1a) but poor predictive performance in the gain domain (Figure 1b).

Analysis of the distribution of fitted parameters (see Figure S2 in SM) suggests that in the gain domain, the best fit of BEAST often reflects a maximal attempt to account for deviations from maximization ($\beta \approx 0$), under the original constraint that the difference between EVs receives considerable weight (50% weight in the original BEAST). AdaBEAST relaxes this extreme constraint (as detailed in the methods section) and improves the prediction accuracy. Linear mixed-effects modeling (see Table S3 in the SM) showed that the difference in MSEs between AdaBEAST and BEAST is significant in both the gain domain, $\beta = -0.0892, t_{(5,536)} = -37.6, p < .001$, and the mixed domain $\beta = -0.0084, t_{(6,841)} = -5.478, p < .001$. This improvement of AdaBEAST can be attributed to the relaxation of the stringent constraints in the original BEAST, providing it with more context

[5] Our prediction for the unknown person's choice in task t assumes he or she is similar to one of the people who are in-sample and will thus behave similarly to that person also in task t. Yet, because the unknown person is equally likely to be similar to any of the individuals in the sample, the best prediction for the unknown person is the average prediction in that task across the sample (or the prediction of an "average new person"). Note this approach does not ignore potential individual differences. Rather, it uses the only available source of heterogeneity for this prediction task, the heterogeneity in the observed sample.

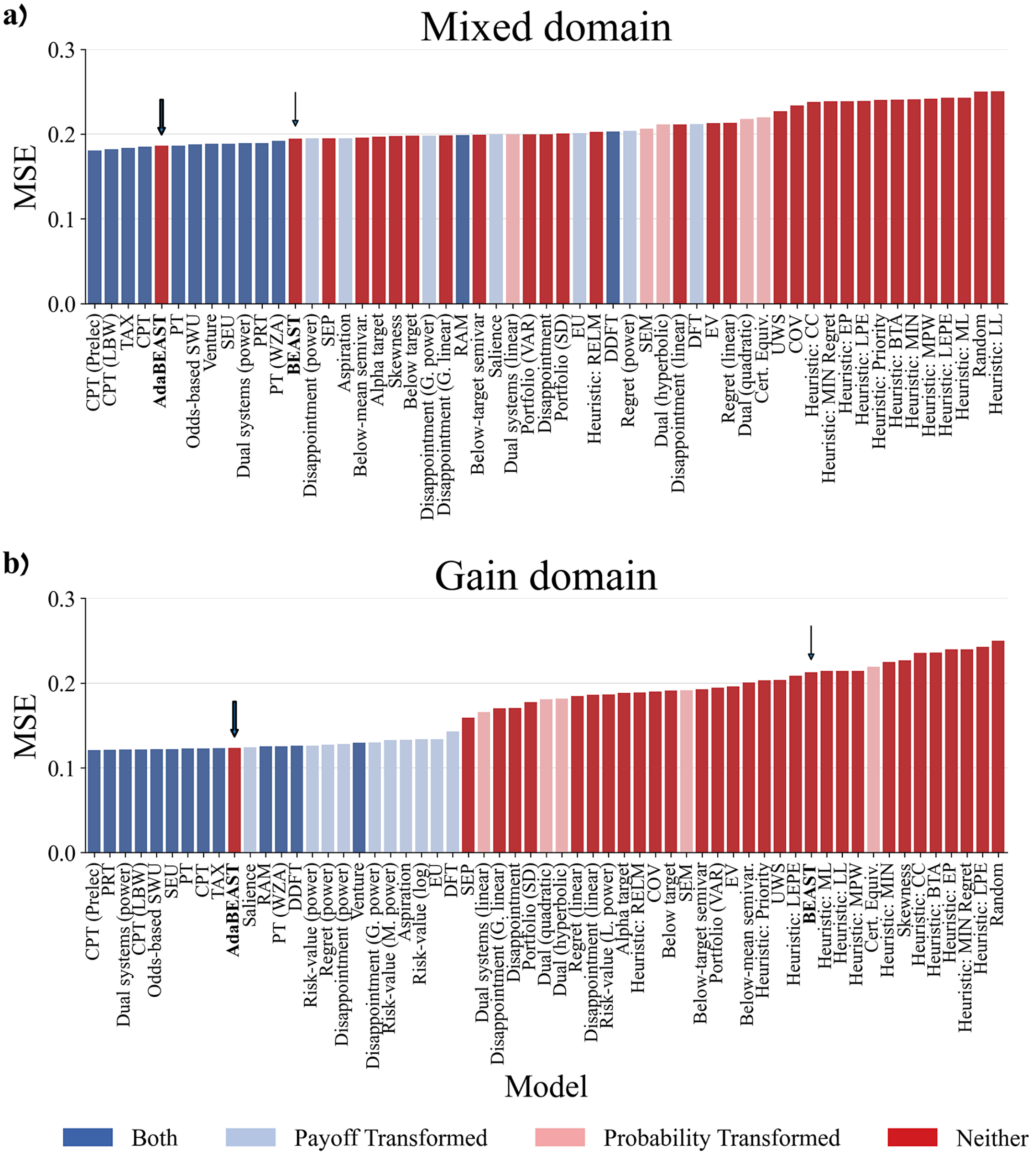

Figure 1. *Average prediction error (MSE) for in-sample individuals, for the (a) mixed domain and (b) gain domain. Bar colors indicate the usage of nonlinear payoff and probability transformations by each of the models. Arrows mark the relative ranking of BEAST (thin arrow) and AdaBEAST (thick arrow).*

adaptability compared to the original BEAST. For example, the distributions of fitted parameters (Figure S3) for AdaBEAST show that often the weights given to the different biased sampling tools are very different than one another, an aspect that cannot be accounted for in the original BEAST.[6]

[6] Moreover, a t-test reveals that the weights given to the unbiased tool are higher in the mixed $[M = .517, SD = .259]$ domain compared to the gain domain $\left[M = .368, SD = .291; T_{(5,981.994)} = 21.772, p < .001, Cohen's\ d = 0.546\right]$. Because relying on many mental draws through the unbiased tool is similar to giving large weight to the EV, these results help explain why the predictions made by the original BEAST (which implied significant weight to the EVs) are far more accurate in the mixed domain datasets compared to the gain domain datasets.

https://doi.org/10.1017/jdm.2024.34 Published online by Cambridge University Press

Overall, AdaBEAST is ranked 5th amongst the behavioral models in the mixed domain and 10th amongst the models in the gain domain (Figure 1, thick arrow). Statistical comparison of AdaBEAST with CPT that uses Prelec (1998) functions, the best behavioral model in this analysis, shows that CPT predicts significantly better than AdaBEAST in the mixed domain, $\beta = 0.0054, t_{(6,842)} = 3.129$, $p = .002$, but the difference in the gain domain is only marginally significant, $\beta = 0.0029, t_{(5,532)} = 1.712, p = .087$.

4.2. Psychological mechanisms

A major focus in the analysis done by (He et al., 2022) was on the exploration of the assumed psychological mechanisms embedded in successful predictive models. The assumption was that if an assumed mechanism consistently appears in the models that predict best, then it is likely an essential mechanism for useful predictions of choice behavior, possibly as it reflects an actual human choice mechanism. In their analysis, HAB observed a clear pattern concerning two specific psychological mechanisms: nonlinear payoff transformation and nonlinear probability transformation. Specifically, they found that all top-performing models integrate both payoff and probability transformations while those at the lower spectrum of performance generally exclude them. This pattern is evident in Figure 1. These results might suggest that for a model to attain top-tier predictive performance of choice under risk, the integration of both of these mechanisms is crucial.

However, this clear pattern is challenged when the nonanalytic BEAST models are included in the comparative assessment. While these models do not assume nonlinear transformations of either payoff or probability, AdaBEAST exhibits strong performance and is ranked among the top models. Intriguingly, AdaBEAST stands as the only model to achieve such top performance without assuming these mechanisms, thereby putting in question the previous pattern that these transformations are indispensable for accurate prediction.

To gain further insight into the success of BEAST, we repeated this qualitative analysis, this time focusing on the psychological mechanisms BEAST incorporates (see Figure S3 in SM). Intriguingly, we found that BEAST and AdaBEAST are the only models among those examined that assume both a regret and a sampling mechanism, pointing to a possible reason that may help explain why AdaBEAST is the only model that performs well despite not assuming either payoff or probability transformations.

Specifically, models that include regret but not sampling normally go over all the possible states of the world, compute the expected regret in each state of the world, and then compute a weighted average of these. BEAST, in contrast, incorporates regret within a sampling framework. In each sample, a single state is “realized” and is incorporated into the decision. Hence, it is quite plausible that not all states of the world will be considered. Consequently, in models with regret but no sampling, it is possible that high regret in a single state of the world will result in an extreme prediction for choice: all decisions are influenced by all states of the world. In BEAST, behavior that does not consider even extreme regret in some states is plausible and expected.

4.3. Dataset-specific analysis

To discern the specific settings in which BEAST performs well or poorly, we analyzed the model’s effectiveness across various individual datasets (see Figure 2). Our findings showed that the poor performance of the original model in the gain domain was largely driven by datasets published by Stewart et al. (2015; 2016). The sets of choice tasks in these studies were specifically designed to elicit within-individual-context effects and aimed to demonstrate how careful task design can alter decision-making behavior in predicted ways. Conversely, most other datasets (e.g., Rieskamp, 2008) derive from studies that mostly incorporated choice tasks that were randomly selected from a large space of tasks, and arguably provide a less biased framework to evaluate models more broadly. Our examination yielded an intriguing revelation: AdaBEAST outperformed all 58 behavioral models in

https://doi.org/10.1017/jdm.2024.34 Published online by Cambridge University Press

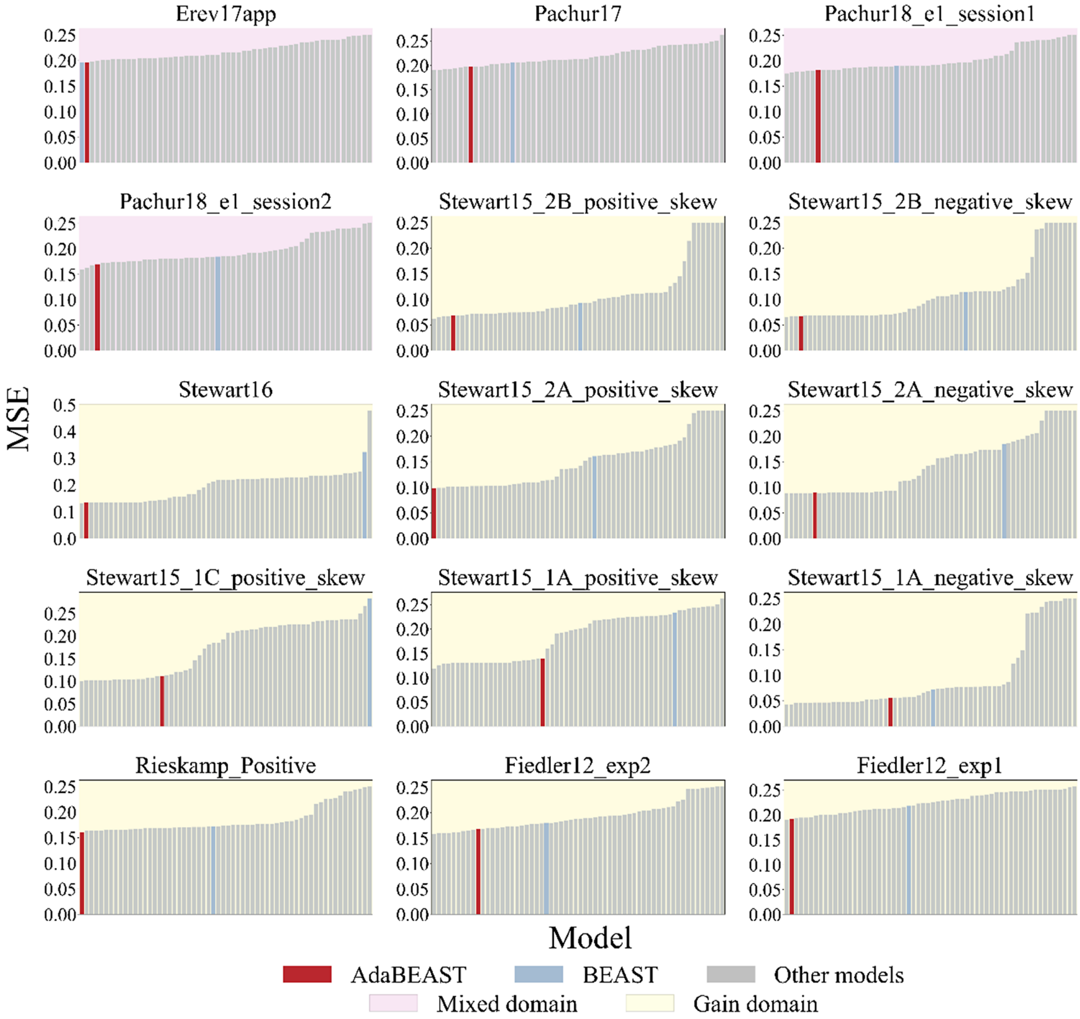

Figure 2. *Prediction error (MSE) for in-sample individuals, by dataset. Each bar represents the prediction error for one of the models, with BEAST and AdaBEAST highlighted. Names of the datasets follow (He et al., 2022), and background colors correspond to the domain of choice tasks.*

HAB's comparison in the only two datasets (Erev et al., 2017; Rieskamp, 2008) that exclusively involved randomly selected choice tasks.

Overall, the nonanalytic BEAST models seem to display a bi-modal performance profile. They tend to falter in cases where choice tasks are systematically generated to elicit idiosyncratic context effects but excel in cases where choice tasks are elicited randomly to cover some large space. This trend underscores the models' potential aptitude for predicting decisions in broader settings.

4.4. Predicting out-of-sample individual decision makers

One of the goals in comparing the 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 real-world environments (e.g., Moisan & Gonzalez, 2017; Raifer et al., 2022). However, in the study

https://doi.org/10.1017/jdm.2024.34 Published online by Cambridge University Press

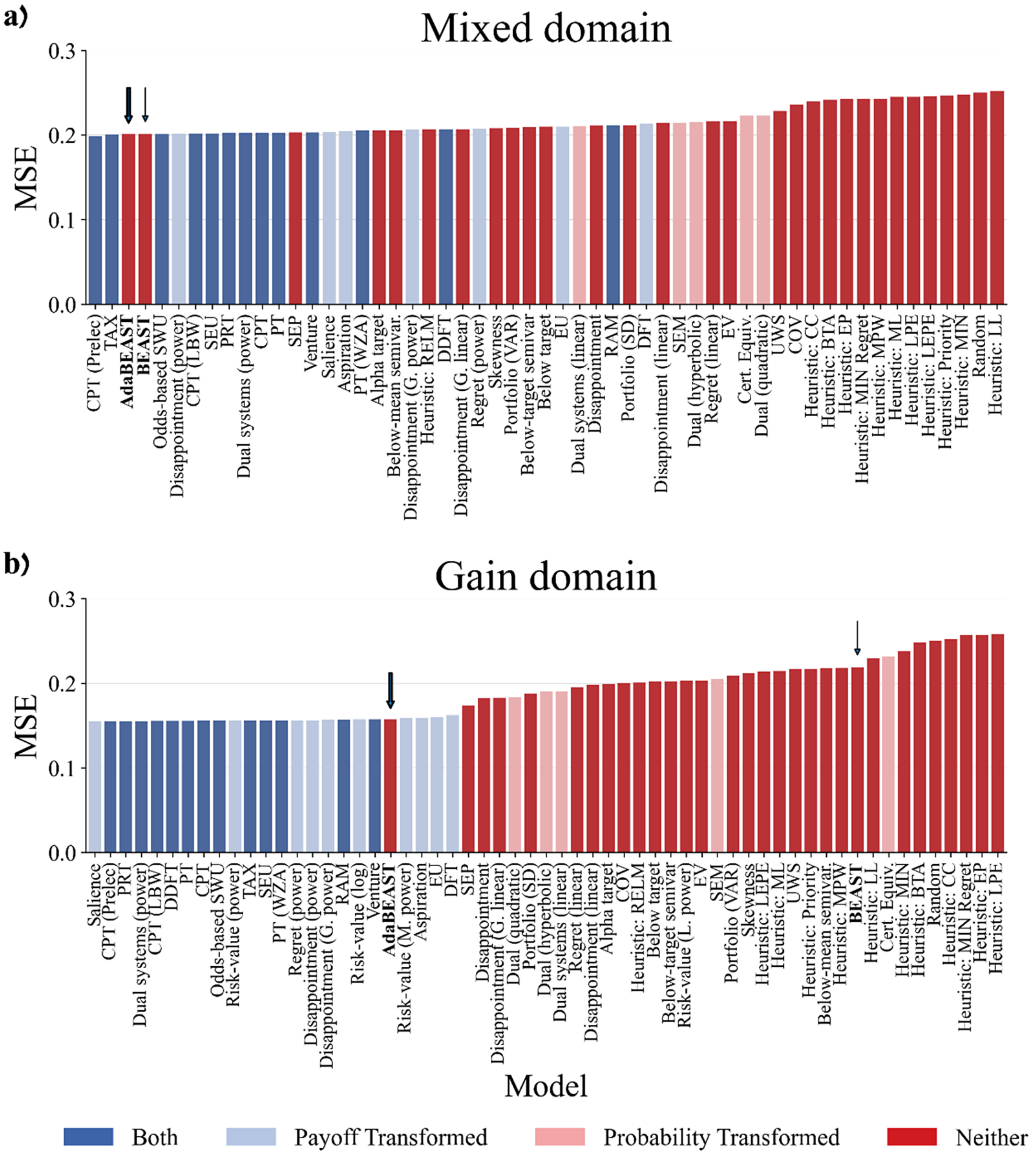

Figure 3. *Average prediction error (MSE) for out-of-sample individuals in the (a) mixed domain and (b) gain domain. Bar colors indicate the usage of nonlinear payoff and probability transformations by each of the models. Arrows mark the relative ranking of BEAST (thin arrow) and AdaBEAST (thick arrow).*

conducted by (He et al., 2022), models were trained and tested on the same sample of participants (although predicting behavior in choice tasks on which models were not trained on). This raises the question of the generalizability of the results to new samples or the population at large. Hence, our subsequent analysis focuses on assessing the predictive power of these models for "unknown", out-of-sample individuals (see methods).

Upon extending our analytical focus to assess predictive accuracy for out-of-sample individuals, the performance landscape showed notable changes (Figure 3; see also Table S4 in the SM). Our analysis yielded not only a remarkable enhancement for the original BEAST model but also a notable change in the rank of AdaBEAST. Specifically, BEAST has now moved up in ranking significantly and, together with AdaBEAST, is among the top three models on HAB's data in the mixed domain (Figure 3a). A linear mixed-effects model with random factors for participants and cross-validation fold of the dataset shows that the performance of both AdaBEAST, $\beta = 0.0016, t_{(6,841)} = 1.147, p = .251$, and the original BEAST model, $\beta = 0.0014, t_{(6,841)} = 0.986, p = .324$, does not differ from that of the most accurate model in the mixed domain (CPT with Prelec functions). Similarly, the difference between AdaBEAST and the most predictive model in the gain domain under this analysis (Salience; Bordalo et al., 2012) was also not significant, $\beta = 0.0021, t_{(5,528)} = 1.455, p = .146$.

5. Discussion

The current research is motivated by inconsistent results observed in prior large-scale decision-making model comparisons and aims to understand the underlying reasons for these differences. While BEAST emerged as the leading predictive model in two CPCs (Erev et al., 2017; Plonsky et al., 2024), a recent analysis by (He et al., 2022) identified a variant of CPT as superior. Our research examines several key methodological differences between these studies that may explain the divergent results. First, in the CPCs, competition participants could submit any model they wished, whereas in HAB the set of competitor models was limited to those adhering to specific criteria set by the authors, and incidentally did not include nonanalytic models. Second, the CPCs featured a broader range of choice tasks, including decision-making under multi-outcome gambles, under ambiguity, and from experience, whereas HAB focused exclusively on decision-making under risk with up to two outcome gambles. Third, models in the CPCs were evaluated based on their predictive accuracy in choice tasks randomly sampled from a large space, whereas in HAB models were evaluated on mostly systematically hand-crafted choice tasks. Lastly, models in the CPCs were required to predict the choice rates of new samples of decision makers, whereas, in HAB, models were required to predict the behavior of individuals already "familiar" to them from the training data. Here, through multiple analyses, we aim to deepen our understanding of the usefulness and predictive efficacy of different models of decisions under risk and offer insights for future studies.

While the CPCs did not apply strict criteria for participation, welcoming all models (including CPT), (He et al., 2022) chose to exclude nonanalytic models because of their complex fitting process. This is an understandable and legitimate criterion since the necessary computational and time resources required for fitting nonanalytic models can be very demanding. However, the exclusion of nonanalytic models may hinder our understanding and can lead to overestimated implications about the psychological mechanisms that help predict decision-making behavior. In their paper, (He et al., 2022) marks nonlinear payoff and probability transformations as essential mechanisms for predictive performance. For example, they write: "Subjective payoff and subjective probability transformation mechanisms stood out as key mechanisms for improving predictive performance in risky choice." (p. 3656). Yet, our analysis demonstrates that accurate predictions of risky choices are possible without relying on nonlinear payoff and probability transformations. BEAST, and more so AdaBEAST predict choice well despite not including these transformations, relying instead on mechanisms like sampling and regret.

The success of our models that combine sampling and regret raises the question of how these models differ from other models that involve either of these mechanisms. According to (He et al., 2022), none of the other evaluated models includes both sampling and regret, making BEAST's unique combination of these mechanisms potentially key to its strong performance. Specifically, unlike some other sampling models (e.g., PRT, Viscusi, 1989), BEAST assumes the mental sampling process is a property of a choice task, rather than the choice options, allowing for context-dependent decisions. Unlike other models that include regret (e.g., SEP, Mellers et al., 1999), BEAST's context dependence is also influenced by the sampled state of the world. Most regret-based models assume decisions are influenced by all possible states, which may lead to extreme predictions due to high regret in a single state. BEAST, however, mentally "realizes" only one state per sample, and as a result, behavior that overlooks even extreme regret in some states is plausible and expected.

When considering datasets primarily comprising randomly selected choice tasks, rather than choice tasks manually crafted by the researchers, the implications of our study become increasingly relevant. While most of the datasets in our analysis included only systematically crafted choice tasks, others also incorporated randomly generated tasks. Analysis of specific datasets revealed that while BEAST tends to perform poorly in contrived, context-specific scenarios, it has marked proficiency in randomly sampled environments; indeed, its adaptable version outperforms all other models in the only two datasets that contain strictly randomly sampled tasks. This divergence in performance foregrounds BEAST's enhanced aptitude for predictive fidelity in settings that potentially contain a broader coverage of the spectrum of possible tasks. The success of our models under such conditions aligns with the original intent for which BEAST was developed. The model was specifically designed to capture a broad spectrum of phenomena in human choice behavior and predict human decision-making in wide sets of environments (Erev et al., 2017).

Despite the success of the nonanalytic models in specific datasets with exclusively randomly sampled choice tasks, our analysis still showed that in the original analysis as conducted by (He et al., 2022), and includes all datasets combined, CPT performs better. One possible reason for this gap may be the fact that CPT is a highly flexible model that when fitted on some choice data by an individual in some context can accurately capture many of that individual-context idiosyncratic interactions well. Indeed, in a recent study, Fudenberg et al. (2023) have asserted that CPT is so flexible 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., 2023, p. 21). In contrast, BEAST was designed with the intent to predict behaviors of new unknown individuals and is congruently far less flexible. To deal with BEAST's rigidity issues, we developed AdaBEAST which allows for increased adaptation across contexts, without fundamentally changing the underlying theory and main underlying assumptions. Indeed, the results showed that AdaBEAST performs better than BEAST in both the mixed domain and the gain domain in which AdaBEAST's prediction power was statistically indistinguishable from that of CPT. However, the results may also suggest that AdaBEAST remains less flexible than models like CPT and therefore does not predict the choices of familiar individuals as well.

Accurate prediction models of human choice can be highly useful in many practical applications. In some cases, like when gauging a patient's adherence to treatment, developing a personalized prediction model for a specific individual is the appropriate approach. But in many other cases, prediction models are most useful when they can predict the behavior of unknown decision makers. This is particularly true when making policy decisions when broad population insights are required, for example for assessing the population response to a planned sugary drink tax or pricing strategies in public transportation. In light of this, in our work, we found it beneficial to also assess the predictive capabilities and generalizability of the models presented in (He et al., 2022), as well as BEAST and AdaBEAST, when applied to new unknown individuals. Indeed, we found that under this analysis the remaining gap between AdaBEAST and CPT (or Salience) is eliminated. These findings accentuate the importance of BEAST's foundational design and its robustness across varied participants.

To predict unknown individuals, we essentially create a prediction for an average new person drawn from the population implied by the observed sample. It may thus be argued that the relative additional success of our nonanalytic models in this task could reflect the lower sensitivity of the models to individual differences. It is probably true that when a model's predictions in a task are not very sensitive to the values of the individual-level parameters its prediction for an average person in the population (based on a sample) would be less noisy and thus better when applied to new individuals. Yet, note that a relatively low (vs. high) sensitivity of the predictions to the choice of parameters is not necessarily evidence against the model's validity. That a model is highly flexible and allows any behavior given different parameters (like CPT; Fudenberg et al., 2023) does not mean that it is necessarily a "proper" model.

https://doi.org/10.1017/jdm.2024.34 Published online by Cambridge University Press

The aforementioned analysis compared the prediction capabilities of the models for new decision-makers in new choice tasks, but it still involved the prediction of behavior in the same experiment and context. That is, to the extent that choice behavior in one task might be influenced by the other tasks that people face in the same experiment (e.g., Ert & Erev, 2013; Schneider et al., 2016; Stewart et al., 2015), training data of models in this analysis involves access to contextual features that may impact behavior within an experiment but will be irrelevant outside of it. Going forward, aligning with methodologies akin to the CPCs, it would be insightful to train models on datasets of specific experiments and participant groups, followed by prediction for entirely new participant groups in new experiments. Such an approach would further our understanding of model generalizability, bringing both theoretical clarity and practical applicability to the fore.

The present study underscores the importance of considering a range of models, including those that may be perceived as more difficult to estimate, as it can add valuable insights into the underlying mechanisms of human behavior. The relative success of BEAST challenges the research community to venture beyond traditional strategies when building models to achieve even better results.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/jdm.2024.34.

Data availability statement. The data used in this study is publicly available from the OSF website: https://osf.io/ca6bn/.

Funding statement. Small parts of this article were previously reported in Agassi & Plonsky (2023). Ori Plonsky acknowledges support from the Israel Science Foundation (grant no. 2390/22).

References

Agassi, O. D., & Plonsky, O. (2023) 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*, 45. Sydney, Australia.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*(5), 961–981.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*, 22–36. https://doi.org/10.2307/1909829

Bordalo, P., Gennaioli, N., & Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics*, *127*(3), 1243–1285. https://doi.org/10.1093/qje/qjs018

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. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44. Toronto, Canada.

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

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*(4), 369–409. https://doi.org/10.1037/rev0000062

Erev, I., Ert, E., Plonsky, O., & Roth, Y. (2023). Contradictory deviations from maximization: Environment-specific biases, or reflections of basic properties of human learning? *Psychological Review*, *130*(3), 640–676. https://doi.org/10.1037/rev0000415

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., & Marx, A. (2023). Humans as intuitive classifiers. *Frontiers in Psychology*, *13*, 1041737.

Erev, I., & Roth, A. E. (2014). Maximization, learning, and economic behavior. *PNAS*, *111*(Suppl. 3), 10818–10825. https://doi.org/10.1073/pnas.1402846111

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.

Fiedler, S., & Glöckner, A. (2012). The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in Psychology*, *3*, 1–18. https://doi.org/10.3389/fpsyg.2012.00335

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy*, *56*, 279–304. https://doi.org/10.1086/256692

Fudenberg, D., Gao, W., & Liang, A. (2023). How flexible is that functional form? Quantifying the restrictiveness of theories. *The Review of economics and statistics*, 1–50. https://doi.org/10.1162/rest_a_01401

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

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review*, *114*(3), 678–703.

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy*, *60*, 151–158. https://doi.org/10.1086/257177

Mellers, B., Schwartz, A., & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General*, *128*(3), 332–345.

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., Mata, R., & Hertwig, R. (2017). Who dares, who errs? Disentangling cognitive and motivational roots of age differences in decisions under risk. *Psychological Science*, *28*(4), 504–518. https://doi.org/10.1177/0956797616687729

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

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., . . . & Erev, I. (2024). Predicting human decisions with behavioral theories and machine learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.1904.06866

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

Plonsky, O., & Erev, I. (2021b). To predict human choice, consider the context. *Trends in Cognitive Sciences*, *25*(10), 819–820.

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

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

Samuelson, P. A. (1963). Risk and uncertainty-a fallacy of large numbers. *Scientia*, *98*, 108–113.

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110–114. https://doi.org/10.2307/3002019

Savage, L. J. (1954). *The foundations of statistics*. (2nd ed). Wiley.

Schneider, S., Kauffman, S., & Ranieri, A. (2016). The effects of surrounding positive and negative experiences on risk taking. *Judgment and Decision Making*, *11*(5), 424–440.

Skinner, B. F. (1953). *Science and human behavior*. Free Press.

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

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

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

Viscusi W. K. (1989) Prospective reference theory: Toward an explanation of the paradoxes. *Journal of Risk and Uncertainty*, *2*(3), 235–263.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Zhu, J. Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719–748.

Cite this article: Agassi, O. D. and Plonsky, O. (2024). Beyond analytic bounds: Re-evaluating predictive power in risky decision models. *Judgment and Decision Making*, e35. https://doi.org/10.1017/jdm.2024.34