Decision

Using Machine Learning to Create an Adaptable, Scalable, and Interpretable Behavioral Model

Vered Shoshan, Tamir Hazan, and Ori Plonsky

Online First Publication, August 22, 2024. https://dx.doi.org/10.1037/dec0000247

AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

ISSN: 2325-9965

https://doi.org/10.1037/dec0000247

Using Machine Learning to Create an Adaptable, Scalable, and Interpretable Behavioral Model

Vered Shoshan, Tamir Hazan, and Ori Plonsky

Faculty of Data and Decision Sciences, Technion, Israel Institute of Technology

In this study, we introduce Adaptable Scalable Best Estimate and Sampling Tools (AS-BEAST), an interpretable model of human decision-making under uncertainty, that fuses the foundational principles of BEAST, a behavioral model grounded in psychological theory, with the capabilities of machine learning techniques. Our strategy involves mathematically formalizing BEAST as a differentiable function and representing it in a computational graph. This approach facilitates the learning of model parameters using automatic differentiation and gradient descent. AS-BEAST scales to larger data sets and adapts to new data more efficiently, while preserving the psychological interpretability of the original model. Evaluation of AS-BEAST on the largest publicly accessible data set of human choice under uncertainty shows that it predicts choice at state-of-the-art levels, similar to those of less interpretable deep neural networks and better than several benchmarks, including the original BEAST model. Importantly, AS-BEAST provides interpretable explanations for choice behavior, leading to the extraction of novel psychological insights from the data. This research demonstrates the potential of machine learning techniques to enhance the scalability and adaptability of models rooted in psychological theory, without compromising their interpretability or insight generation capabilities.

Keywords: behavioral economics, human decision making, interpretability, cognitive models

Supplemental materials: https://doi.org/10.1037/dec0000247.supp

Imagine two scenarios: a sure gain of 1,000 or a 50% chance of gaining 2,000 (and nothing otherwise). Both have equal expected values (EVs), yet people usually opt for the certain gain, avoiding the risk of getting nothing. This is an example of a classical choice task that provides insight into human decision-making under risk and uncertainty, a critical research area in economics and behavioral sciences with impacts on societal and engineering issues, such as public policy design, understanding consumer behavior, and creating artificial intelligence agents that mimic human choice (Amir et al., 2005; Apel et al., 2022; Barberis, 2013; Bhargava & Loewenstein, 2015; Chetty, 2015; List, 2004; Reisch & Zhao, 2017; Rosenfeld & Kraus, 2018; Russell, 2010).

Historically, decision making was thought to be rational, maximizing expected utility (EU). However, behavioral research repeatedly shows that humans often deviate from rationality. There have been numerous attempts to model these deviations, with prospect theory (PT; Kahneman & Tversky, 1979) the most prominent example.

Ori Plonsky https://orcid.org/0000-0002-1619-4482

This research was financially supported by the Ministry of Science and Technology of Israel.

Vered Shoshan played a lead role in formal analysis, investigation, visualization, and writing–original draft and a supporting role in methodology and writing–review and editing. Tamir Hazan played a supporting role in conceptualization, funding acquisition, supervision, and writing–original draft. Ori Plonsky played a lead role in conceptualization, methodology, project administration, funding acquisition, supervision, and writing–review and editing and a supporting role in formal analysis, investigation, and writing–original draft.

Correspondence concerning this article should be addressed to Ori Plonsky, Faculty of Data and Decision Sciences, Technion, Israel Institute of Technology, Haifa 3200003, Israel. Email: plonsky@technion.ac.il

Yet, finding a precise predictive model remains a major challenge: Many of the documented deviations from rationality contradict one another, and it is difficult to predict when and where these different deviations occur.

One highly successful model is Best Estimate and Sampling Tools (BEAST; Erev et al., 2017), a psychological theory-based behavioral model that uses simulations of assumed human mental processes to approximate population-wide choice tendencies. As Erev et al. (2017) showed, BEAST not only captures 14 major (e.g., loss aversion) and sometimes contradictory (e.g., both overweighting and underweighting of rare events) deviations from maximization, but also predicts choice surprisingly accurately, better than both other behavioral models and data-driven models. Despite its success, BEAST struggles with scalability and adaptability, limiting its utility for new data (He et al., 2022; Peterson et al., 2021).

To address this, machine learning (ML) models utilizing BEAST's quantitative predictions and its core insights as features have been proven effective in predicting human decision making (Plonsky et al., 2017, 2024). However, ML models' contribution is limited when it comes to providing explanations, interpreting behavior, and inferring causality—key areas where theory-based behavioral models can be beneficial.

In this study, we develop Adaptable Scalable BEAST (AS-BEAST), a behavioral model that utilizes ML approaches but maintains high interpretability. We first transform the main underpinnings of the (simulation-based) model BEAST into a differentiable mathematical function with a small set of behaviorally interpretable parameters. This then allows us to use common approaches in ML, like automatic differentiation (AD) and gradient descent for efficient parameter estimation, despite the relative complexity of the underlying model. Because this approach vastly improves the efficiency of model training, it also allows us to (a) relax multiple restricting assumptions embedded in the original BEAST thus making the model more adaptable to new data sets, and (b) scale the model to much larger data sets than before. Crucially, our approach preserves the interpretability of the original model. We evaluate AS-BEAST on the largest human choice task data set available, demonstrating that it predicts choice out of sample better than several baselines, including BEAST itself, and at a similar level to the performance of highly flexible but less interpretable deep neural networks. Importantly, AS-BEAST provides interpretable explanations for choice behavior, leading to new psychological insights from the data.

Related Work

Interest in ML applications in economics has increased, particularly in behavioral economics (Altman et al., 2006; Bourgin et al., 2019; Camerer, 2018; Erev et al., 2017; Fudenberg et al., 2022; Fudenberg & Liang, 2019; Hartford et al., 2016; Kleinberg et al., 2017; Peysakhovich & Naecker, 2017; Plonsky et al., 2017, 2024) and econometrics (Mullainathan & Spiess, 2017). Efforts have largely focused on prediction improvement, with only few studies leveraging ML approaches to gain new psychological insights (Fudenberg et al., 2022; Fudenberg & Liang, 2019; Kleinberg et al., 2017; Peysakhovich & Naecker, 2017). One notable contribution was made by Hartford et al. (2016), who used a neural network to model behavior in economic games and gain insights by introducing inductive bias. Yet, extracting behavioral insights from ML models remains challenging.

Our work extends Peterson et al. (2021), who tested, using a large data set (Bourgin et al., 2019), various prediction methods for decision making, ranging from psychological theory-only models (including BEAST and prospect theory) to fully data-driven neural networks. In their work, they examined the predictive accuracy of BEAST, without retraining it on their data due to its scalability problems. Despite this limitation, they found that the untrained BEAST was impressively accurate, performing similarly to the best behavioral models adapted using neural networks and trained on their data. The best model was mixture of theories (MoT), a network learning weights for different theories per choice task. Although challenging due to BEAST's simulation-based nature, we similarly aim to optimize learning through gradient-based optimization.

Decision-Making Under Uncertainty

Human decision-making under uncertainty is often studied using the task of choice among M gambles, each $m \in M$ defined with N outcomes, $\{x_i^m\}_{i=1}^N$ and their respective probabilities $\{p_i^m\}_{i=1}^N$. Like previous works (Bourgin et al., 2019; Erev et al., 2017; Hartford et al., 2016; Plonsky et al., 2017, 2024), our article focuses on a model predicting the population proportion choosing each option in a given choice task, such as choosing between two gambles.

Behavioral Models of Decision Making

Classical decision-making models assume a utility function U that measures the value of each outcome to a decision maker. Here, the rational choice maximizes the EU: the aggregate of utilities multiplied by their corresponding probabilities (Von Neumann & Morgenstern, 1947). However, these models fail to capture frequent deviations from rationality, or "behavioral anomalies," in human choices. To better portray these findings, researchers have developed dozens of descriptive decision-making models (He et al., 2022), with one of the most successful examples being BEAST (Erev et al., 2017).

BEAST proposes a decision-making process involving noisy and sometimes biased mental sampling of potential outcomes. To forecast the proportion of the population choosing each of the M gambles, BEAST simulates this mental sampling process for many agents and calculates the proportion choosing each gamble. In the present context, $M = 2$, and we refer to the gambles as "A" and "B." Each simulated agent then chooses B over A if and only if:

$$\Delta\mathrm{BEV} + \Delta\mathrm{ST} + e > 0, \tag{1}$$

where $\Delta\mathrm{BEV} = \mathrm{BEV}_B - \mathrm{BEV}_A$, and BEV_m is the best estimate of the EV of m; e is an error term: a random number normally distributed with mean zero; and $\Delta\mathrm{ST} = \mathrm{ST}_B - \mathrm{ST}_A$, where ST_m ("sampling tools") is the average of few values, each sampled from one of four possible distributions: the gamble's unbiased outcome distribution, and three biased transformations of the it that represent assumed psychological biases. These biases are a tendency to assume the worst ("pessimism"), a tendency to treat all outcomes as if they are equally likely ("uniform"), and a tendency to focus only on the outcomes' sign, ignoring magnitudes ("sign"). Finally, BEAST assumes a somewhat different process when one of the gambles stochastically dominates the other, and specifically zero noise.[1]

In addition to this theory-grounded process, the original BEAST design includes auxiliary assumptions like equal weights for ΔBEV and ΔST and equal weight to each of the biased sampling tools. These assumptions were introduced to expedite training and avoid overfitting on the small data set based on which BEAST was developed. We achieve much faster training under much larger data sets, and so our adaptation of BEAST omits these assumptions for increased flexibility and potential new behavioral insights.

Flexibility Versus Interpretability in Human Decision-Making Under Uncertainty

Behavioral models based on psychological or economic theory are often designed for interpretability. Yet, they often fall short in capturing the full range of observed behavior. These models, including EU and prospect theory, offer inadequate results even when adapted using neural networks (Peterson et al., 2021). BEAST, in contrast, stands out for its accuracy. For example, models based on BEAST won two choice prediction competitions (Erev et al., 2017; Plonsky et al., 2024) and, without any retraining, can predict new data at a similar level to established behavioral models that were both enhanced using neural networks and trained on that new data (Peterson et al., 2021). However, BEAST is not scalable. Generating each of its predictions requires running many time consuming simulations, and, more importantly, because these simulations make its output not differentiable with respect to the parameters, training BEAST requires brute-force approaches which are often computationally prohibitive. Specifically, in its original implementation, BEAST's parameters were fitted using a grid search over a discrete set of predetermined and somewhat

[1] A stochastically dominates B if for any outcome x, $P(A \geq x) \geq P(B \geq x)$, with strict inequality for some x.

arbitrary possible values. Furthermore, as mentioned, BEAST's original implementation includes multiple auxiliary assumptions that reduce the number of parameters that are needed to be fit to the data. The addition of these restrictive assumptions to BEAST allowed training of the model to a medium-size data set despite the necessity to use computationally demanding simulations in the training process. However, these assumptions also greatly reduce the model's flexibility and adaptability and fitting BEAST to larger data sets remains computationally challenging even when these assumptions are maintained.

ML models in human decision making are scalable, easily adaptable, and often highly accurate. However, their "black-box" nature makes them far less interpretable. Consequently, it is hard to analyze and understand their predictions, particularly in terms of the underlying causes of the human's behavior. Models of context-free and artificial decision-making under uncertainty, like models for human choice between gambles, are mainly aimed to guide scientific thinking and enhance understanding of human decision making in more natural settings. Hence, this lack of interpretability significantly reduces the usefulness of these models. To overcome these challenges, we propose adapting an interpretable psychological theory-based model like BEAST, known for its predictive prowess, to incorporate the scalability and flexibility of ML.

AS-BEAST

While models of ML focus on prediction, theories in behavioral decision making focus on explanation (Hofman et al., 2021). Our work aims to merge the adaptability of ML algorithms with psychological theory-based behavioral models for improved predictions and new behavioral insights. Given its proven success as a behavioral model (Erev et al., 2017; Peterson et al., 2021), we specifically chose to adapt BEAST. Our approach first transforms BEAST from a simulation-based to a mathematical, differentiable model for easier training on new data. This involves representing the expected decision BEAST predicts as a differentiable function of the model's parameters and the properties of the choice task. However, due to the model's complexity, deriving analytical expressions for the partial derivatives of each model parameter is challenging. To address this, we represent the model as a differentiable computational graph—essentially a network of mathematical operations that produces the model's predictions as a function of its (highly interpretable) parameters—which allows us to use a process of AD. AD is a common technique in ML to compute gradients of complex functions with high precision. Unlike traditional methods that approximate these gradients or require manual calculation, AD navigates through the computational graph, tracking and applying the chain rule at each operation, thereby computing the exact gradients needed for optimization. We then employ gradient-based optimization, an iterative method where the model parameters are gradually adjusted in the direction that minimizes the loss (the error between the model's predictions and the data). This approach, common in ML applications that train complex models, not only simplifies the training process but also enhances its scalability.

Converting BEAST From a Simulation Model to a Mathematical Model

Before introducing AS-BEAST, it is useful to explain the mechanics of the original simulation-based model BEAST.

BEAST

BEAST's prediction for choosing Gamble B over A, $\text{Pred}_{\text{BEAST}}$, is the average of many simulated decisions, each of which makes a choice as per Equation 1. In practice, with n simulations:

$$\text{Pred}_{\text{BEAST}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\Delta\text{BEV}_i + \Delta\text{ST}_i + e_i > 0}, \qquad (2)$$

such that $\mathbb{1}$ is an indicator function.

In the context of choice between gambles with full information, in which computation of the gamble's EV is possible:[2]

[2] BEAST can also predict human choice under ambiguity, when EVs cannot be directly computed. Here and in all benchmark works, predictions are limited to the classical full information setting. In full information, we mean that the gambles are fully known to the decision maker at time of choice.

$$\mathrm{BEV}_{mi} = \mathrm{EV}_m = \sum_{j=1}^{N} x_j^m p_j^m. \quad (3)$$

Furthermore, the terms in Equations (1) and (2), are defined such that e_i is a normally distributed error term and:

$$\mathrm{ST}_{mi} = \frac{\sum_{j=1}^{\kappa_i} \mathrm{sample}_i(T_j(m))}{\kappa_i}, \kappa_i \sim U(1, K), \quad (4)$$

$$T_j = \begin{cases} \text{unbiased} & \text{w.p } p_{\text{unb}} \\ \text{uniform} & \text{w.p } p_{\text{unif}} \\ \text{pessimism} & \text{w.p } p_{\text{pess}} \\ \text{sign} & \text{w.p } p_{\text{sign}} \end{cases}, \quad (5)$$

where $\mathrm{sample}_i(T_j(m))$ is an outcome sampled using sampling tool T_j applied on gamble m, κ_i is the number of sampling tools drawn in simulation i, and K is the maximal number of sampling tools to draw, a model parameter (note draws are independent so the same tool can be selected more than once). Importantly, the theory underlying BEAST restricts K to be *small*, so that the model will reflect "reliance on small samples," an assumption that was found to be highly useful for explaining and predicting decision-making under uncertainty (Erev et al., 2023; Plonsky et al., 2015). Furthermore, p_{unb}, p_{unif}, p_{pess}, p_{sign} are the probabilities of using the sampling tools Unbiased, Uniform, Pessimism, and Sign, respectively.[3] These sampling tools are transformations on the outcome distribution of the gamble.[4] Specifically, for gamble $m = \{x_i^m\}_{i=1}^N, \{p_i^m\}_{i=1}^N$:

$$\text{unbiased}(m) = \{x_i^m\}_{i=1}^N, \{p_i^m\}_{i=1}^N, \quad (6)$$

$$\text{uniform}(m) = \{x_i^m\}_{i=1}^N, \{\frac{1}{N}\}_{i=1}^N, \quad (7)$$

$$\text{pessimism}(m) = \min(\{x_i^m\}_{i=1}^N), 1, \quad (8)$$

$$\text{sign}(m) = \{\text{sign}(x_i^m)\}_{i=1}^N, \{p_i^m\}_{i=1}^N. \quad (9)$$

For illustration of the process in BEAST, consider a choice between Gamble A that gives "1,000 with certainty" and Gamble B that gives "50% chance of gaining 2,000, and nothing otherwise." Because both gambles have the same EV, $\Delta\mathrm{BEV} = 0$. BEAST thus predicts that the choice rate will be determined by the output of the sampling tools alone (plus noise). In each simulation, some combination of sampling tools is chosen. For example, assume that in simulation i, $\kappa_i = 2$ and the two sampling tools drawn are Unbiased and Pessimism. Further, assume that for Gamble B, the Unbiased tool (unbiased draw from the distribution) produces the outcome 0. Because this tool will always produce the outcome 1,000 from Gamble A and because the Pessimism tool produces the worst outcome from each gamble, we get: $\mathrm{ST}_{Ai} = \frac{1000+1000}{2} = 1000, \mathrm{ST}_{Bi} = \frac{0+0}{2} = 0$, and thus $\Delta\mathrm{ST} = -1{,}000$. This indicates that in simulation i BEAST will likely predict choice of Gamble A (but the actual prediction will also depend on the drawn value of the error term). Alternatively, if for Gamble B the Unbiased tool produces the outcome 2,000, then $\mathrm{ST}_{Bi} = \frac{0+2000}{2} = 1000$, and thus $\Delta\mathrm{ST} = 0$, implying indifference that will be resolved only by the error term.

AS-BEAST

As previously mentioned, and as can be seen in the equations above, the predictions of BEAST are not differentiable. Our primary goal in development of AS-BEAST is to approximate $\mathrm{Pred}_{\mathrm{BEAST}}$ using a differentiable function. Moreover, as discussed above, BEAST's implementation includes several restrictive auxiliary (i.e., not theory-driven) assumptions. Another goal is to therefore modify some of these assumptions to improve the model's adaptability. AS-BEAST modifies BEAST in the following manner.

To obtain a differentiable prediction, we utilized the expectation of the decision, as the average of many simulated decisions converges to the expected decision:

$$\lim_{n\to\infty} \mathrm{Pred}_{\mathrm{BEAST}} = \mathbb{E}[\mathbb{1}_{g+e>0}] = P[\mathbb{1}_{g+e>0} = 1] = P[e > -g], \quad (10)$$

[3] In the original BEAST, $p_{\text{unif}} = p_{\text{pess}} = p_{\text{sign}}$ and are determined by a function of the number of trials with feedback and one free parameter—see Erev et al. (2017). In AS-BEAST, we remove this restriction and consider them free parameters. In both models, p_{unb} complements to sum of probabilities to 1.

[4] BEAST assumes a somewhat different implementation of the sampling tools Unbiased and Uniform before versus after decision makers get feedback on their choice. Moreover, sampling tool Pessimism is not triggered when all outcomes are in the loss domain—see Erev et al. (2017), for details. In AS-BEAST, we maintain all these assumptions and implement the sampling tools accordingly.

where $g = \Delta BEV + \Delta ST$. To approximate this probability, we modify the error term of the model using the logistic function (see Bowling et al., 2009):

$$P[e > -g] \approx \frac{1}{1 + \exp^{-b \cdot g}}, \tag{11}$$

where b is a parameter that determines the curvature of the logistic function sigmoid. Larger values of b imply a lower weight for the noise e relative to the input g.

We next find a term for the expectation of $\frac{1}{1+\exp^{-b \cdot g}}$ as a function of the model's parameters. From Equation (3), it follows that in the full information case on which we focus here, ΔBEV is constant, and the expectation of $\frac{1}{1+\exp^{-b \cdot g}}$ depends only on ΔST. Therefore:

$$\mathbb{E}_{\Delta ST}\left[\frac{1}{1 + \exp^{-b \cdot g}}\right] = \sum_{\Delta ST} P(\Delta ST) \frac{1}{1 + \exp^{-b \cdot g}}. \tag{12}$$

Hence, it is necessary to go over all possible scenarios that could lead to different values of ΔST and find their corresponding probabilities. As can be derived from Equations (4) and (5), ΔST depends on the random variables κ and $T_1, \ldots T_\kappa$ and on the output of the sampling process using these $T_1, \ldots, T_\kappa$. This implies that given values for κ and $T_1, \ldots, T_\kappa$, we can derive all possible values of ΔST for each gamble. Furthermore, for any given choice task, it is also possible to directly calculate $P(\Delta ST|\kappa, T_1, \ldots, T_\kappa)$ (see Supplemental Material).

To illustrate, we return to the example task above ("1,000 with certainty" vs. "50% for 2,000 or nothing"). Here, with $\kappa = 2$, $T_1 =$ Unbiased, and $T_2 = Pessimsim$, it is possible to get $\Delta ST = -1{,}000$ (as before), and it is also possible to get $\Delta ST = 0$ (which happens when the output of the Unbiased tool applied to Gamble B equals 2,000), but no other value. Further, $P(\Delta ST = -1000|\kappa = 2, T_1 = \text{Unbiased}, T_2 = Pessimism) = P(\Delta ST = 0|\kappa = 2, T_1 = \text{Unbiased}, T_2 = Pessimism) = 0.5$. Similarly, in any given choice task, we can go over all possible combinations of κ, $T_1, \ldots, T_\kappa$ and derive all possible values of both ΔST and $P(\Delta ST|\kappa, T_1, \ldots, T_k)$. In our model, this is done in the data preprocessing stage (see below) since these values depend only on the features of the choice task and not on the model's parameters.

A natural way to derive an expression for resolving Equation 12 is then to condition over all possible combinations of κ, $T_1, \ldots, T_\kappa$. In particular, using the law of total probability, and because $T_1, \ldots, T_k$ are all independent in each other:

(see Equation 13 below)

As mentioned, $P(\Delta ST|\kappa, T_1, \ldots, T_k)$ are all directly calculable. Furthermore, according to Equation (4), κ is a uniform random variable, that is, $P(\kappa) = \frac{1}{K}$. Moreover, according to Equation (5), $P(T_1) \ldots P(T_\kappa) \in \{p_{unb}, p_{uni}, p_{pes}, p_{sign}\}$. Thus, we can get an expression of $P(\Delta ST)$ as a function of K, p_{unb}, p_{uni}, p_{pes}, and p_{sign}, which are all parameters that can be learned from the data. Although K, the maximal number of sampling tools used, can be learned (and indeed is a free parameter in the original BEAST), as it increases, the number of possible sampling tool combinations increases exponentially, which greatly impacts computational complexity. Since in the original implementation, K was estimated to equal 3, we chose, in the present study, to fix $K = 3$. Note this also restricts the model to reflect "reliance on small samples," a theoretical restriction embedded in BEAST.

Finally, BEAST's original implementation assumes, quite arbitrarily, that $p_{unif} = p_{pess} = p_{sign}$, and that the weight given to the difference between the (best estimates of the) EVs ΔBEV and the difference between the averages of potentially biased mental samples ΔST are equal (see Equation 1). To increase the model's adaptability, in AS-BEAST, we remove these arbitrary restrictions. Specifically, we assume that the weight of ΔBEV relative to the weight of ΔST is a free parameter to be learned, w.

$$P(\Delta ST) = \sum_{\kappa=1}^{K} P(\kappa) \sum_{T_1} P(T_1) \ldots \sum_{T_\kappa} P(T_\kappa) P(\Delta ST|\kappa, T_1, \ldots, T_\kappa). \tag{13}$$

Combining Equations (12) and (13) we get:

$$\begin{aligned}\text{Pred}_{\text{AS-BEAST}} &= \mathbb{E}_{\Delta\text{ST}}[d(\Delta\text{ST})] \\ &= \sum_{\Delta ST} P(\Delta\text{ST})d(\Delta\text{ST}),\end{aligned} \quad (14)$$

$$d(\Delta\text{ST}) = \frac{1}{1+\exp^{-b(w\cdot\Delta\text{BEV}+\Delta\text{ST})}}, \quad (15)$$

and $P(\Delta\text{ST})$ is a function of p_{unb}, p_{uni}, p_{pes}, p_{sign}, as described in Equation (13).

Hence, the predictions of AS-BEAST depend on six parameters: a weight w for ΔBEV in the decision relative to ΔST, the b parameter of the logistic function approximating the non-differentiable decision, and the four sampling tools probabilities p_{unb}, p_{uni}, p_{pes}, p_{sign}. Note that these four probabilities must sum to 1 and hence the model has only 5 degrees of freedom.

Preprocessing

The preprocessing stage, illustrated on the left hand-side of Figure 1, involves computing, for each choice task, ΔBEV (#1 in the figure) as per Equation 3, all possible values of $P(\Delta\text{ST}|\kappa, T_1, \ldots, T_\kappa)$ (#5) as per Equation (13), and all possible values of ΔST (#4). For the latter, we create (a) 84 vectors of possible ΔSTs, with each vector containing the potential values of ΔST conditional on a specific combination of $T_1, \ldots, T_\kappa$ applied to gambles A and B (#2 in the figure), and (b) 84 vectors of possible $P(\Delta\text{ST}|\kappa, T_1, \ldots, T_\kappa)$, with each vector storing the probabilities to sample outcomes using that combination of sampling tools $T_1, \ldots, T_\kappa$ (#3 in the figure).[5] Note that in both cases, the number of elements in each vector equals the number of possible ΔST values under a particular combination of sampling tools $T_1, \ldots, T_\kappa$. For instance, in our running example ("1,000 with certainty" vs. "50% for 2,000 or nothing"), the vectors corresponding to the combination of sampling tools $\{T_1 =$ Unbiased, and $T_2 = Pessimsim\}$ include two elements each; specifically the vectors are $(-1{,}000,\ 0)$ and $(0.5,\ 0.5)$ for the ΔST and $P(\Delta\text{ST}|\cdot)$ respectively. Of course, the vectors may have different lengths under a different combination of sampling tools in this task and/or in other tasks.

Differentiable Computational Graph

To train the model, we implement a computational graph, as shown on the right hand-side of Figure 1. The graph receives three inputs from the preprocessing stage for each choice task: ΔBEV, the set of vectors holding all possible ΔSTs, and the set of vectors holding all possible $P(\Delta\text{ST}|\kappa, T_1, \ldots, T_\kappa)$. It computes the term $d(\Delta\text{ST})$ according to Equation 14 (#6 in the figure) for each element in each of the ΔST vectors as well as the probability of each possible term $P(\Delta\text{ST})$ for each element in each of the $P(\Delta\text{ST}|\kappa, T_1, \ldots, T_\kappa)$ vectors according to Equation 13 (#7 in the figure). These computations depend on the learned parameters b, w, p_{unb}, p_{uni}, p_{pes}, p_{sign} (#11). The graph then computes $\text{Pred}_{\text{AS-BEAST}}$ as per Equation 14 (#8) and calculates the loss (#10) between the ground truth (#9) and its prediction. It then performs a backward pass to compute gradients by which updating parameters would minimize the loss and updates the parameters accordingly. This updating is done iteratively to optimize the free parameters based on the training data.

Experiment

Method

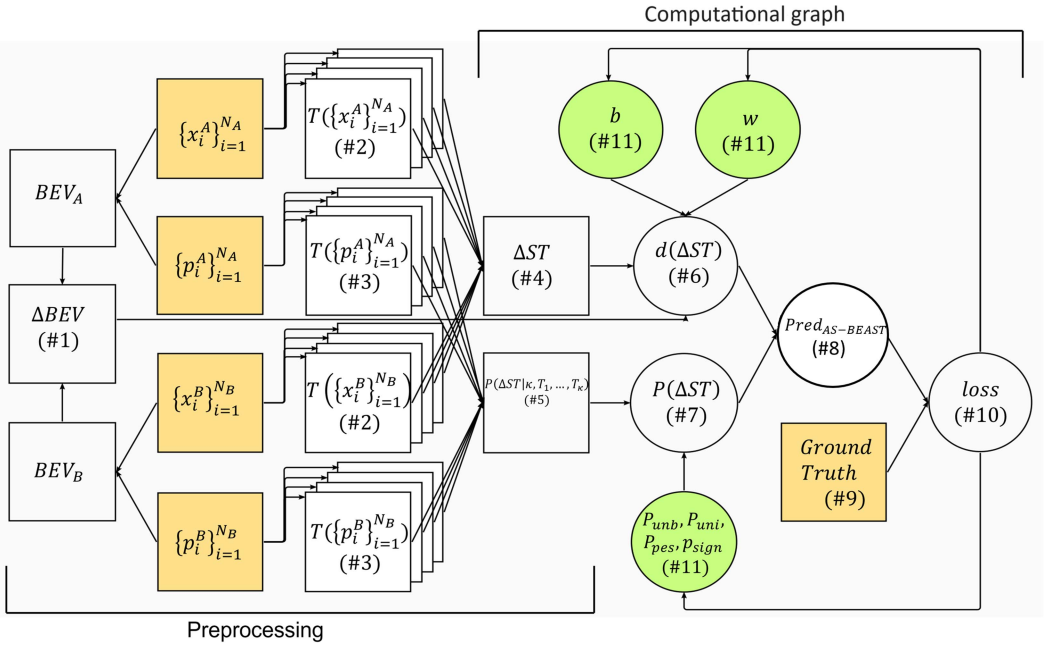
Choices13k Data Set

Our main experiment involves training and evaluating AS-BEAST on the largest public data set of human risky choice behavior, Choices13k (Bourgin et al., 2019). This data set includes 13,006 different choice tasks (selecting between Gambles A and B) in each of which human decision makers made choices for five trials in sequence. In the tasks, one of the options may include up to 10 outcomes and the other option may include up to two outcomes. Further details concerning the data set and tasks are given in Bourgin et al. (2019) and Peterson et al. (2021).

Of the 13 k tasks, as in Peterson et al. (2021), we focus on the 9,831 fully described binary choice tasks in which full feedback concerning both the obtained and the forgone payoffs is provided after each of the five trials. Although

[5] In each combination, there are $\kappa \in \{1, 2, 3\}$ sampling tools, that are each one of four different tools, so there are $\sum_{i=1}^{3} 4^i = 84$ such combinations.

Figure 1
Illustration of AS-BEAST Training Procedure



Note. Squares represent parts that are prepossessed, with orange squares the original task's input (see "Preprocessing"). Circles represent the differentiable computational graph where the gradients are computed, with green circles for the learned behavioral parameters (see "Differentiable computational graph"). AS-BEAST = Adaptable Scalable Best Estimate and Sampling Tools. See the online article for the color version of this figure.

feedback on previous choices usually has an impact on people's choices, the published data are not given per trial. Rather, it includes, for each choice task, the mean aggregate choice rate for each gamble over all five trials. To preserve its generality, AS-BEAST learns different parameter values and generates separate predictions for the choice rate in each trial of a sequence of choices (so it can be applied to data sets with different lengths of sequences, including one-shot choices). Hence, here, in each choice task, we average AS-BEAST's five trial-level predictions and compare the result with the available data. Congruently, we report the average of the parameter values learned across the five trials of the sequence.

Model Setup

We randomly split the data such that 80% of the available data (the train set) was used for model training, the other 20% of the data (the test-set) was used for model evaluation. This process was repeated 10 times, and we report the average test-set mean squared error (MSE) ± 1 SD. The best training parameters were found to be: data set size for number of epochs, batch size of 10, stochastic gradient descent as the optimizer, and a learning rate of 0.09. Initial values for p_{unb}, p_{uni}, p_{pes}, p_{sign} were their EVs in the original BEAST. Initial values for b and w were 0.2 and 0.3. Code to reproduce the results is in the Supplemental Material.

Tasks With Stochastically Dominant Options

The original BEAST model presumes a different process when one gamble stochastically dominates another. In a similar vein, other works have also explicitly altered predictive models of choice under risk and uncertainty when they deal with such tasks (Peterson et al., 2021; Plonsky et al., 2018). Following these works and the logic in the original BEAST, we decided to treat tasks with stochastic dominance separately from tasks without stochastic dominance. Our different

treatment of tasks with stochastic dominance takes one of two forms: fixed rate or model tuning. The fixed rate method was employed by Peterson et al. (2021) in their most successful model for the current data (MoT, see below). It includes simply predicting a fixed choice rate for the dominating option in all relevant tasks of the test-set, based on the average choice rate of such options in all relevant tasks in the train set. While this method is effective for prediction purposes, it cannot explain *how* decision makers behave differently in tasks with and without stochastic dominance. The model tuning method produces more interpretable results. It includes training the model separately only on (train set) tasks without stochastic dominance and then using the resulting parameter values found in this training as the initial values for training the model only on the subset of choice tasks with stochastic dominance. Under this method, the model training produces different sets of parameters for tasks with and tasks without stochastic dominance and since the parameters of AS-BEAST are highly interpretable, it gives rise to explanations for the nature of the difference in behavior between the two types of tasks. Specifically, BEAST originally assumes less noisy choices in tasks with stochastic dominance. In AS-BEAST, this would translate a higher value for the *b* parameter. Hence, we can use the model tuning method to examine the assumption embedded in BEAST, in addition to investigating if these tasks are different in other ways than nondominant tasks.

Benchmark Models

We compare the predictive performance of AS-BEAST to four interpretable behavioral models: neural EU, neural prospect theory, neural cumulative prospect theory, and the original BEAST. The first three of these were developed and trained by Peterson et al. (2021) as neural networks implementing classical behavioral models (EU, prospect theory, and cumulative prospect theory, respectively), essentially optimizing the functional forms of these models based on the training data. Note that, as explained by Peterson et al. (2021), these models all belong to a class of models in which potential gambles are valued independently of their alternatives, a strong theoretical restriction which is possibly too restrictive (Peterson et al., 2021; Plonsky & Erev, 2021).This restriction also implies that these models cannot explicitly distinguish between tasks with and tasks without dominated options, since to identify a dominance relation, both gambles must be considered jointly. That is, under these models, it is not theoretically feasible to treat tasks with and tasks without dominance distinctly. Therefore, to the extant that a distinct treatment of such tasks is useful (e.g., as assumed in BEAST), the predictive performance of these models is likely to suffer.

The fourth benchmark model we consider, the original BEAST, is not trained to the Choices13k data set. Its parameter values are taken from the original implementation (Erev et al., 2017). Indeed a main motivation for the development of AS-BEAST is that training of BEAST in this large data set is computationally challenging.

The final benchmark model we compare AS-BEAST to is MoT, the currently best performing published model for the current data. As mentioned, MoT uses a fixed rate method for its predictions in tasks with dominance. For tasks without dominance, MoT uses a neural network to learn weights that should be given to the predictions of each of two interpretable behavioral models (also implemented as neural networks) per data point. Since the neural network assigning weights is not interpretable, it is a priori unclear what it is in a choice task that makes people behave as if they give more weight to one model over the other. This property, we argue, makes MoT a less interpretable model than other models we consider here. The performance of all benchmark models except the original BEAST is taken directly from Peterson et al. (2021).

Segmenting Tasks Without Dominance to Subgroups

Because of AS-BEAST's high level of interpretability, it is possible to use the model to probe how specific properties of the choice task impact choice behavior. Thanks to size and diversity of the Choices13k data set, we can partition the tasks to several mutually exclusive subgroups, tune the model separately on each subgroup, and analyze the parameters learned within each subgroup for behavioral insights. This segmentation involved only tasks without stochastic dominance. We explored two segmentation approaches for the tasks without dominance: manual, theory-driven, segmentation based on behavioral decision-making literature observations, and an automatic,

data-driven, segmentation using deep clustering via K-Means algorithm on an autoencoder's latent representation.

Manual (Theory-Driven) Segmentation. Traditional studies of choice between gambles considered mostly gambles with up to two possible outcomes. Yet, research suggests that choice behavior can alter when the number of outcomes in a gamble increases (Bernheim & Sprenger, 2020; Venkatraman et al., 2014). Building on this, we segmented the (no-dominance) choice tasks to four groups, according to the number of outcomes in their gambles: Tasks with a gamble of one outcome versus a gamble with two ("1vs2"); two outcomes versus two ("2vs2"); one outcome versus three or more ("1vsM"); and two outcomes versus three or more ("2vsM").

Automatic (Data-Driven) Segmentation. To segment We incorporated psychological features from Plonsky et al. (2017), in addition to the original BEAST prediction and the unsegmented (model tuning) AS-BEAST prediction, into the task's features. The psychological features are derived directly from the model BEAST and reflect the elements BEAST assumes the decision process is sensitive to. Prior work using ML to predict human behavior demonstrated the necessity of these features for better ML prediction with smaller data sets. We employed an autoencoder-based implementation (https://github.com/xuyxu/Deep-Clustering-Network) based on Yang et al. (2017) to learn a latent representation that is "easy to cluster" with the K-Means algorithm and set $K = 4$.

Results

AS-BEAST implementation and training takes significantly less time than BEAST. For instance, where BEAST took over a day to generate predictions on Choices13k without retraining (Peterson et al., 2021), AS-BEAST produces results in minutes, demonstrating improved scalability.

Prediction Accuracy

Table 1 shows the predictive accuracy of (the different variants of) AS-BEAST in Choices13k in comparison with the benchmark models, in terms of the test-set *MSE* between observed and predicted choice rates. The results show that AS-BEAST's *MSE* is nearly half of that of all other similarly interpretable models we examine, including the original BEAST. This marked improvement is not merely due to different treatment of tasks with and tasks without stochastic dominance: The *MSE* of the model for tasks without stochastic dominance is 0.0123, only slightly worse than the accuracy of the model over all tasks combined, and much lower than that of these benchmark models. AS-BEAST's corresponding accuracy of the modal choice (i.e., proportion of tasks in which the model and data agree on which gamble is more commonly selected) is 82.8%. Importantly, the predictive accuracy of AS-BEAST with fixed rate prediction for the tasks involving dominance is remarkably similar to that of MoT, which uses the same method for predicting tasks involving dominance. This implies that both models perform equally well in both tasks with and tasks without dominance. Yet, as mentioned, we argue that AS-BEAST is more interpretable than MoT.

Whereas some of the large improvement of performance from BEAST to AS-BEAST can naturally be attributed to the fact that BEAST has not been trained on this data, it is likely that some of it involves AS-BEAST increased adaptability. The top rows in Table 2 present the parameter values that the training process has found for AS-BEAST, with the corresponding EVs of these parameters as implied by the original BEAST.

Table 1
Test MSE on Choices13k Data Set

Model	*MSE*
Neural expected utility	0.0217
Neural prospect theory	0.0204
Neural cumulative prospect theory	0.0210
Original BEAST[a]	0.0214
Mixture of theories (MoT)[b]	0.0113
AS-BEAST	
Fixed-rate for tasks with dominance	0.0113 ± 0.0002
Model tuning for tasks with dominance	0.0118 ± 0.0003
AS-BEAST with segmentation[c]	
Manual (theory-driven)	0.0106 ± 0.0002
Automatic (data-driven)	0.0115 ± 0.0003

Note. MSEs when predicting all test-set tasks (both tasks with and tasks without dominant gambles). *MSE* = mean squared error; AS-BEAST = Adaptable Scalable Best Estimate and Sampling Tools.
[a] Model not trained on Choices13k data. [b] Model less interpretable than other models in the table. [c] With the fixed rate method for predicting tasks with dominance.

Table 2
Values of Parameters Learned by AS-BEAST

Model	b	w	p_{unb}	p_{uni}	p_{pes}	p_{sign}
BEAST	NA	1	.64	.12	.12	.12
AS-BEAST, on Choice13k						
Tasks without dominance	0.35	0.16	.59	.17	.04	.20
Tasks with dominance (model tuning)	1.31	0	.37	.29	.05	.30
AS-BEAST with segmentation, on Choice13k						
1vs2	0.40	0.29	.53	.25	.03	.19
2vs2	0.23	0.24	.59	.14	.03	.24
1vsM	0.73	0	.54	.24	.03	.19
2vsM	0.26	0.12	.61	.13	.03	.23
AS-BEAST on synthBEAST	0.43	0.84	.65	.13	.1	.12

Note. AS-BEAST = Adaptable Scalable Best Estimate and Sampling Tools; NA = not applicable; 1vs2 = one outcome versus a gamble with two; 2vs2 = two outcomes versus two; 1vsM = one outcome versus three or more; 2vsM = two outcomes versus three or more; synthBEAST = synthetically created data generated using the model BEAST (see "Validation of Parameter Values").

The results highlight that two of the arbitrary assumptions embedded in BEAST to save free parameters are refuted in AS-BEAST that drops them. Specifically, BEAST assumes that (a) the weight given to the difference between EVs, ΔBEV, is equal to the weight given to the difference between the outputs of the mental sampling process, ΔST, (i.e., it enforces $w = 1$) and (b) that the probability of using each of the biased sampling tools is equal (i.e., it enforces $p_{uni} = p_{pes} = p_{sign}$). As the learned values in AS-BEAST imply, neither of these arbitrary assumptions are held in the current data. We return to analyze these learned values in the "Interpreting Behavior" section below.

Although the main purpose of segmentation of tasks (that do not involve dominance) is to discover patterns of behavior that the model implies differ between different types of tasks (see below), it is still useful to examine how the segmentation impacts the prediction accuracy of the model. Segmentation that improves the model's accuracy is worthy of further examination. The results (bottom lines in Table 1) suggest that manual (theory-driven) segmentation, inspired by previous behavioral studies that suggested the number of outcomes in a gamble affects behavior, outperformed the automatic (data-driven) segmentation. This implies that the theory-driven segmentation constructs more useful groups of tasks in which the model's parameters are shared.[6] Moreover, AS-BEAST with this type of segmentation (and fixed rate prediction for dominant tasks) is more accurate ($MSE = 0.0106$) than all other models presented for this data thus far in terms of *MSE*, although not in terms of the accuracy in predicting the modal choice (83.3%) with MoT reported results slightly more accurate (84.2%). Manual segmentation also improves the accuracy of the model when using the model tuning approach to predict tasks with dominance ($MSE = 0.0112$), making it the second most accurate model for these data (not shown in the table).

Interpreting Behavior

A key advantage of AS-BEAST over other highly predictive models for this data (like MoT) is that it has a small number of highly interpretable parameters whose values are learned and can be used for behavioral insights. The learned values, shown in Table 2, indicate several patterns.

First, in AS-BEAST, the estimated pessimism sampling tool usage probability, p_{pes}, is close to zero and considerably lower than in the original BEAST. This finding aligns with the fact that in Choices13k, participants could not lose money (had they chosen a losing gamble, their actual payoff was 0), contrasting with BEAST's original data set (Erev et al., 2017). This choice of experimental design makes pessimism less sensible here. Given this very low estimated probability, we trained a version of AS-BEAST that removes the availability of the Pessimism sampling tool entirely.

[6] The Supplemental Material demonstrates that the data-driven clustering resulted with very different clusters of tasks than the theory-based clustering.

The results show that the predictive performance of this variant of the model is virtually identical to that of the full model, indicating that for the current data, assuming people have a tendency for pessimism is not very useful. Notably, when we repeated this exercise by removing the Sign sampling tool that, according to the estimated parameter values, is used in nonnegligble proportion of the time, the accuracy of the model was heavily impacted ($MSE = 0.0149$). We take this as evidence that the parameter values that are learned from the data have a meaningful and useful interpretation.

Second, we can compare the parameter values for tasks with dominance and tasks without them. Indeed, this comparison is a major reason to use the "model tuning" method for prediction of tasks with dominance. While a fixed rate prediction appears to yield better results, tuning the model on dominant tasks allows for greater insight into what exactly changes in behavior between tasks with and tasks without dominance. The results show that the most striking difference between the two types of tasks is a substantially larger b parameter value, indicating smaller noise, in tasks with dominance. This aligns with the original BEAST model's assumption, as explained above. Another interesting finding concerning the parameter values in tasks with dominance is zero weight to the difference between the gamble's EVs. This is likely because in tasks with dominance one option is strictly better than the other, regardless of the magnitude of the difference between their EVs, which therefore has no impact on the choice rate.

Next, we move to AS-BEAST with (theory-driven) segmentation, in order to derive insights from the way the model applies in different subsets of tasks. As shown in Table 2, the w parameter, signifying the relative weight of EV differences, is notably smaller in tasks where one option includes more than two outcomes ("1vsM" and "2vsM"), than in tasks that do not include multiple outcomes. This intuitive result may suggest that multiple outcomes make computation and reliance on EVs more challenging. This finding is consistent with past behavioral decision-making research (Huck & Weizsäcker, 1999; Venkatraman et al., 2014).

Finally, in "1vsM" tasks, the model learned a much higher b parameter value, indicating lower noise and more extreme predictions, as well as zero weight to the difference between EVs (w), indicating significant reliance on the output of the mental sampling process. Furthermore, the model learned a relatively high value for the weight of the Unbiased sampling tool (p_{unb}) and thus appears to predict a particularly high reliance on the output of this tool. Because reliance on a small number of unbiased samples from a distribution implies underweighting of rare events (due to properties of binomial distribution, see e.g., Erev et al., 2023), the learned parameters appear to reflect increased underweighting of rare events in the "1vsM" subset of tasks (relative to other subsets). This is a behavioral insight that, to the best of our knowledge, is new to the literature. To verify this finding holds, we examined the observed correlations between the choice rates of the gambles and the proportion of trials that the chosen gamble can be expected to outperform the unchosen gamble. Higher positive correlations reflect more underweighting of rare events. We found indeed that the correlation is higher in the "1vsM" subset of tasks ($r_{1\text{vm}} = 0.47$) than in the other subsets (rs' range [0.34, 0.42]). Thus, the high level of interpretability of the learned parameters helped us reveal an intriguing behavioral phenomenon that we were not aware of beforehand. We further delved into these apparent differences in behavior between subsets of tasks and found that it is useful to assume that in multioutcome gambles, decision makers sometimes ignore some of the available outcomes, perhaps due to their complexity. This naturally increases behavior that appears as underweighting of rare events in multioutcome gambles, particularly when they are contrasted with a certain payoff. Indeed, incorporating this assumption into the original BEAST model reduced its *MSE* on the Choices13k data set by over 5.3%.

Validation of Parameter Values

To establish that the above discussion that uses the values of the model parameters for interpreting behavior is grounded, we made sure that AS-BEAST can recover known parameter values, and specifically the parameter values used in the original BEAST model. For this purpose, we trained AS-BEAST on a new synthetically created data set, synthBEAST, that was produced using the same method as the synthetic data set in Bourgin et al. (2019). Specifically, we generated choice tasks using the same algorithm used to

[7] Training used the same hyperparameters and initial values as in the main experiment, with the exception of the values for b and w which were set to 0.8 and 1, respectively.

create choice tasks in Choices13k and tagged them using the original BEAST model. We then trained the model on this synthetic data set[7] and examined the values of the learned parameters. The results, shown in Table 2, suggest that AS-BEAST effectively learned parameter values that closely resembled those of the original BEAST model. This suggests that AS-BEAST can accurately reconstruct behavioral patterns and thus interpreting its parameters in different data sets can provide behavioral insights.

Discussion

Our novel approach integrates ML strengths in scalability and flexibility with the interpretability of theory-based behavioral models. We exemplify this through the transformation of the simulation-based BEAST model into an adaptable computational graph that can then be trained and produce predictions with enhanced accuracy and speed. Crucially, the refined model not only retains its theoretical interpretability but also unveils new behavioral phenomena, setting the stage for potential discoveries in future data sets.

Our study is motivated by the observation that the original BEAST model was found to be a surprisingly useful model for prediction of choice under risk and uncertainty (Agassi & Plonsky, 2023; Erev et al., 2017; Peterson et al., 2021; Plonsky et al., 2024). Yet, it is also an extremely hard model to scale to large data sets like Choices13k. Congruently, previous attempts to examine its performance on this data did not train it, but used the parameters fitted to another data set. Notably, training BEAST requires a grid search over a selected space of potential parameter values. The speed of this procedure is highly dependent on the relatively arbitrary presuppositions concerning the discrete set of allowable parameter values that create the grid, and its accuracy depends on the resolution of the search. Moreover, when deriving the predictions of BEAST for each set of possible parameters, the modeler should choose some accuracy level, reflected in the number of simulations that are run. A higher number of simulations increases the accuracy but also the computational demands. AS-BEAST, in contrast, is trained using AD and stochastic gradient descent and therefore does not restrict the values of the parameters to a discrete closed and likely small set. Further, it also aims to directly compute the expected decision of an infinite number of simulations, making accuracy concerns less relevant. In principle, after we created a differentiable version of BEAST, we could have used other, more traditional methods to estimate the parameters. Yet, ML methods are particularly adept at handling complex cost functions, which may feature multiple local minima, and they do not require calculation of second derivatives—a task that is often intractable for sophisticated models. Moreover, the use of AD facilitates efficient computation of derivatives without manual derivation that can be difficult in complex models. Finally, these methods also offer superior scalability and computational efficiency, particularly with large data sets. On the other hand, to train effectively, these methods require large data sets. This could be particularly problematic in the realm of behavioral sciences, where data sets are often relatively small, although this is changing in recent years. Further, these methods can be less precise than more traditional methods. Nevertheless, as we show, using these methods in our context can effectively recover known parameters and provide meaningful values.

Our use of ML approaches differs from regular applications of ML in developing predictive models. Rather than employing ML directly to produce predictions, we make use of common methods used in ML to estimate parameters of a theory-grounded model. While this type of applications is novel in the domain of behavioral decision making, similar approaches has been used recently for estimating parameters in the context of neurophysiological models of neuroimaging data (Griffiths et al., 2022) and of agent-based models of opinion dynamics (Lenti et al., 2024). The main purpose in these types of works does not rely necessarily on achieving superior predictive accuracy but on accurately and efficiently estimating interpretable models. In the current work, we achieve both efficient parameter estimation and state-of-the-art levels of prediction accuracy.

Our strategy of synergizing ML approaches with behavioral models offers wider applicability. For instance, in the "1vsM" outcome problems of the Choices13k data set, we discovered that decision makers behave as if they tend to disregard some outcomes. This behavioral pattern could in principle be modeled using a "dropout layer" common in neural networks, that can be added to the computational graph, with the "dropout probabilities" (the likelihood of an outcome being ignored) learned from the data. This outcome–dropout approach can be employed in any

behavioral model adapted using computational graphs, not just BEAST, by inserting a dropout layer on the gamble outcomes while upholding the standard logic of the behavioral model.

One limitation of AS-BEAST, particularly when contrasted with BEAST, is the fact that it is designed to directly produce a predicted average choice of the population in a choice task, rather than simulate the choices individual decision makers make in that task. In this sense, AS-BEAST is less of a process model than BEAST that makes assumptions about the actual choice process individual decision makers make before averaging them to make a prediction. Moreover, AS-BEAST, as implemented, can only be applied to choice between fully described gambles (i.e., without ambiguity), whereas BEAST is more general. Yet, for the purpose of predicting the population-wide choice rates in the classic decisions under risk setting, AS-BEAST is sufficient.

Conclusion

In our research, we unveiled AS-BEAST, a pioneering model that blends the scalability and adaptability of ML with the interpretability of a psychological theory-based model in predicting decision-making under uncertainty. This was achieved through the incorporation of ML capabilities for gradient-based optimization within a mathematical and more flexible adaptation of the BEAST model.

AS-BEAST's performance was validated on the most extensive publicly available data set on decision-making under uncertainty, where it outperformed strong benchmarks. The model's adaptability was instrumental in enhancing BEAST's prediction accuracy and runtime, but perhaps more crucially, its interpretability yielded fresh behavioral insights.

We anticipate our work will pave the way for an innovative approach that synergizes theory and ML, fostering greater collaboration between domain experts and ML researchers. More specifically, we hope our model encourages deeper engagement between the ML and behavioral science communities, thereby enhancing our ability to predict human decision making.

References

Agassi, O. D., & Plonsky, O. (2023). The importance of non-analytic models in decision making research: An empirical analysis using BEAST. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*. https://escholarship.org/uc/item/3wm9278q

Altman, A., Bercovici-Boden, A., & Tennenholtz, M. (2006). Learning in one-shot strategic form games. In J. Fürnkranz, T. Scheffer, & M. Spiliopoulou (Eds.), *Machine Learning: ECML 2006: 17th European Conference on Machine Learning* (pp. 6–17). Springer. https://doi.org/10.1007/11871842_6

Amir, O., Ariely, D., Cooke, A., Dunning, D., Epley, N., Gneezy, U., Koszegi, B., Lichtenstein, D., Mazar, N., Mullainathan, S., & Prelec, D. (2005). Psychology, behavioral economics, and public policy. *Marketing Letters*, *16*(3), 443–454. https://doi.org/10.1007/s11002-005-5904-2

Apel, R., Erev, I., Reichart, R., & Tennenholtz, M. (2022). Predicting decisions in language based persuasion games. *Journal of Artificial Intelligence Research*, *73*, 1025–1091. https://doi.org/10.1613/jair.1.13510

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, *27*(1), 173–96. https://doi.org/10.1257/jep.27.1.173

Bernheim, B. D., & Sprenger, C. (2020). On the empirical validity of cumulative prospect theory: Experimental evidence of rank-independent probability weighting. *Econometrica*, *88*(4), 1363–1409. https://doi.org/10.3982/ECTA16646

Bhargava, S., & Loewenstein, G. (2015). Behavioral economics and public policy 102: Beyond nudging. *American Economic Review*, *105*(5), 396–401. https://doi.org/10.1257/aer.p20151049

Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J., & Griffiths, T. L. (2019). Cognitive model priors for predicting human decisions. *Proceedings of the 36th International Conference on Machine Learning, PMLR*, *97*, 5133–5141. https://proceedings.mlr.press/v97/peterson19a.html

Bowling, S. R., Khasawneh, M. T., Kaewkuekool, S., & Cho, B. R. (2009). A logistic approximation to the cumulative normal distribution. *Journal of Industrial Engineering and Management*, *2*(1), 114–127. https://doi.org/10.3926/jiem.2009.v2n1.p114-127

Camerer, C. F. (2018). Artificial intelligence and behavioral economics. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 587–608). University of Chicago Press.

Chetty, R. (2015). Behavioral economics and public policy: A pragmatic perspective. *American Economic Review*, *105*(5), 1–33. https://doi.org/10.1257/aer.p20151108

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

Fudenberg, D., Kleinberg, J., Liang, A., & Mullainathan, S. (2022). Measuring the completeness of economic models. *Journal of Political Economy*, *130*(4), 956–990. https://doi.org/10.1086/718371

Fudenberg, D., & Liang, A. (2019). Predicting and understanding initial play. *American Economic Review*, *109*(12), 4112–4141. https://doi.org/10.1257/aer.20180654

Griffiths, J. D., Wang, Z., Ather, S. H., Momi, D., Rich, S., Diaconescu, A., McIntosh, A. R., & Shen, K. (2022). *Deep learning-based parameter estimation for neurophysiological models of neuroimaging data*. bioRxiv. https://doi.org/10.1101/2022.05.19.492664

Hartford, J. S., Wright, J. R., & Leyton-Brown, K. (2016). Deep learning for predicting human strategic behavior . *Proceedings of Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper/2016/hash/7eb3c8be3d411e8ebfab08eba5f49632-Abstract.html

He, L., Analytis, P. P., & Bhatia, S. (2022). The wisdom of model crowds. *Management Science*, *68*(5), 3635–3659. https://doi.org/10.1287/mnsc.2021.4090

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Huck, S., & Weizsäcker, G. (1999). Risk, complexity, and deviations from expected-value maximization: Results of a lottery choice experiment. *Journal of Economic Psychology*, *20*(6), 699–715. https://doi.org/10.1016/S0167-4870(99)00031-8

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

Kleinberg, J., Liang, A., & Mullainathan, S. (2017). The theory is predictive, but is it complete? An application to human perception of randomness. *Proceedings of the 2017 ACM Conference on Economics and Computation*, 125–126. https://doi.org/10.1145/3033274.3084094

Lenti, J., Monti, C., & De Francisci Morales, G. (2024). Likelihood-Based methods improve parameter estimation in opinion dynamics models. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 350–359. https://doi.org/10.1145/3616855.3635785

List, J. A. (2004). Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica*, *72*(2), 615–625. https://doi.org/10.1111/j.1468-0262.2004.00502.x

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

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. https://doi.org/10.1126/science.abe2629

Peysakhovich, A., & Naecker, J. (2017). Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, *133*, 373–384. https://doi.org/10.1016/j.jebo.2016.08.017

Plonsky, O., Apel, R., Erev, I., Ert, E., & Tennenholtz, M. (2018). *When and how can social scientists add value to data scientists? A choice prediction competition for human decision making* [Unpublished manuscript]. Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., Carter, E. C., Cavanagh, J. F., & Erev, I. (2024). *Predicting human decisions with behavioral theories and machine learning*. arXiv. https://doi.org/10.48550/arXiv.1904.06866

Plonsky, O., & Erev, I. (2021). To predict human choice, consider the context. *Trends in Cognitive Sciences*, *25*(10), 819–820. https://doi.org/10.1016/j.tics.2021.07.007

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017). Psychological forest: Predicting human behavior. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.10613

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

Reisch, L. A., & Zhao, M. (2017). Behavioural economics, consumer behaviour and consumer policy: State of the art. *Behavioural Public Policy*, *1*(2), 190–206. https://doi.org/10.1017/bpp.2017.1

Rosenfeld, A., & Kraus, S. (2018). *Predicting human decision-making: From prediction to action*. Springer, Cham. https://doi.org/10.1007/978-3-031-01578-6

Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education.

Venkatraman, V., Payne, J. W., & Huettel, S. A. (2014). An overall probability of winning heuristic for complex risky decisions: Choice and eye fixation evidence. *Organizational Behavior and Human Decision Processes*, *125*(2), 73–87. https://doi.org/10.1016/j.obhdp.2014.06.003

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd rev.). Princeton University Press.

Yang, B., Fu, X., Sidiropoulos, N. D., & Hong, M. (2017). Towards k-means-friendly spaces: Simultaneous deep learning and clustering . *Proceedings of the 34th International Conference on Machine Learning, PMLR*, *70*, 3861–3870. https://proceedings.mlr.press/v70/yang17b.html

Received August 3, 2023
Revision received May 8, 2024
Accepted May 20, 2024 ■