nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02267-6

Predicting human decisions with behavioural theories and machine learning

Received: 17 April 2024

Accepted: 11 June 2025

Published online: 21 July 2025

Check for updates

Ori Plonsky[1], Reut Apel[1], Eyal Ert[2], Moshe Tennenholtz[1], David Bourgin[3], Joshua C. Peterson[4], Daniel Reichman[5], Thomas L. Griffiths[6], Stuart J. Russell[7], Even C. Carter[8], James F. Cavanagh[9] & Ido Erev[1]

Predicting human decisions under risk and uncertainty remains a fundamental challenge across disciplines. Existing models often struggle even in highly stylized tasks like choice between lotteries. Here we introduce BEAST gradient boosting (BEAST-GB), a hybrid model integrating behavioural theory (BEAST) with machine learning. We first present CPC18, a competition for predicting risky choice, in which BEAST-GB won. Then, using two large datasets, we demonstrate that BEAST-GB predicts more accurately than neural networks trained on extensive data and dozens of existing behavioural models. BEAST-GB also generalizes robustly across unseen experimental contexts, surpassing direct empirical generalization, and helps to refine and improve the behavioural theory itself. Our analyses highlight the potential of anchoring predictions on behavioural theory even in data-rich settings and even when the theory alone falters. Our results underscore how integrating machine learning with theoretical frameworks, especially those—like BEAST—designed for prediction, can improve our ability to predict and understand human behaviour.

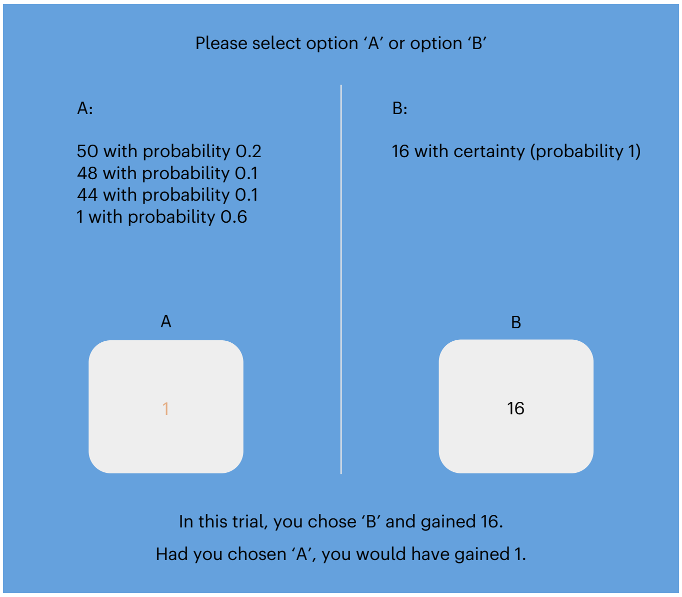
Many human decisions in health, finance, environment and management occur under risk and uncertainty. Understanding and predicting such decisions is a fundamental goal in fields such as economics, psychology, cognitive science and artificial intelligence. Indeed, decision-making under uncertainty has been a central topic of research since Bernoulli's work nearly three centuries ago[1]. Although this research has led to valuable insights and to the development of many behavioural models grounded in empirical phenomena and/or theoretical constraints[2–4], no single model consistently and accurately describes and predicts choices across even the most basic stylized tasks, such as choice between lotteries.

Recent large-scale studies have sought to identify a model capable of such robust prediction[5–7]. In one study[5], a choice prediction competition, researchers submitted models predicting human choice between lotteries, and the models were evaluated on the basis of their predictive accuracy in new held-out data. With the focus on prediction accuracy, one might expect machine learning (ML) tools to excel. Indeed, ML tools have a strong predictive record across domains, including in the prediction of human choice under uncertainty[8–11], and their predictive power is often assumed to provide an upper bound on the possible accuracy of behavioural descriptive models[12–15]. However, the competition and additional analysis have shown that behavioural-theory-free ML performed poorly compared with models incorporating behavioural insights. Chief among these were variants of the behavioural model BEAST (best estimate and sampling tools)[5]. Interestingly, BEAST makes very different assumptions than those assumed by mainstream models such as prospect theory. Whereas most models were designed to clarify interesting deviations from expected utility theory, BEAST was designed to predict behaviour and posits that choices result from a potentially biased mental sampling process and sensitivity to expected values (EVs).

Subsequent studies revealed boundary conditions on BEAST's dominance. Plonsky et al.[16] demonstrated that an ML algorithm using

[1]Technion – Israel Institute of Technology, Haifa, Israel. [2]The Hebrew University of Jerusalem, Jerusalem, Israel. [3]Adobe Research, San Francisco, CA, USA. [4]Boston University, Boston, MA, USA. [5]Worcester Polytechnic Institute, Worcester, MA, USA. [6]Princeton University, Princeton, NJ, USA. [7]University of California, Berkeley, Berkeley, CA, USA. [8]DEVCOM Army Research Laboratory, Adelphi, MD, USA. [9]The University of New Mexico, New Mexico, NM, USA. e-mail: plonsky@technion.ac.il

features derived from the behavioural assumptions of BEAST outperformed BEAST itself—and all other models—on the competition's data. Similarly, Peterson et al.[7] showed that, when deep neural networks are designed to reflect theoretical behavioural assumptions, they can efficiently and accurately predict choice in similar tasks. These findings thus suggest that the hybrid approach, combining ML with behavioural features, can harness the strengths of both, augmenting the predictive power of ML with domain-relevant knowledge. However, Peterson et al.[7,17] also demonstrated that, with sufficiently large datasets, purely data-driven neural networks can very accurately predict risky choice. This suggests that, given enough training samples, behavioural insights may not add much. Consequently, it is unclear which approach best predicts human choice on new data: strictly behavioural models like BEAST, behavioural-theory-free ML trained on ample data, or hybrid models that integrate ML with behavioural theories.



Please select option ‘A’ or option ‘B’

A:

50 with probability 0.2
48 with probability 0.1
44 with probability 0.1
1 with probability 0.6

B:

16 with certainty (probability 1)

A

1

B

16

In this trial, you chose ‘B’ and gained 16.

Had you chosen ‘A’, you would have gained 1.

Fig. 1 | Example decision-making task used in CPC18. Human decision-makers chose between A and B repeatedly for 25 trials. In the first five trials, they received no feedback, and starting trial 6 they received full feedback (on both obtained and forgone outcomes) after each choice. The figure presents the feedback screen in one trial in one of the 270 different choice tasks used in CPC18. The participant in this example chose option B.

Here, we start by presenting the design and results of another choice prediction competition, CPC18, that expanded the space of choice tasks examined in the original competition and explicitly encouraged submissions that involve ML[18]. A key advantage of the competition format is that it reduces the risk of overlooking alternative modelling strategies by inviting many distinct approaches to the same predictive challenge. The winning submission, from five of the current authors (D.B., J.C.P., D.R., T.L.G. and S.J.R.), was a hybrid model called BEAST gradient boosting (BEAST-GB). BEAST-GB combines BEAST's quantitative predictions and features engineered based on the assumptions of BEAST with an extreme gradient boosting (XGB) algorithm[19]. Its success reinforces the idea that combining ML and behavioural logic yields superior predictions of human choice in this domain.

We then proceed to examine the performance of BEAST-GB in two other large datasets, each illuminating different facets of the hybrid approach. First, using the largest public dataset of human choice between lotteries, we test whether BEAST behavioural insights (implemented as features) remain valuable even when the training data are substantially increased. That is, we check if purely data-driven ML can learn the behavioural choice patterns without direct access to domain-specific theoretical logic. We also analyse the differences between the predictions of BEAST and those of BEAST-GB to uncover predictable patterns in choice behaviour that BEAST misses. Second, we use a large meta-dataset recently compiled to compare dozens of decision-making models in a different decision-making task to examine how much value ML can add above and beyond the performance of the behavioural model. Specifically, we check whether, even when BEAST itself predicts poorly, a hybrid leveraging its structure still excels. Together, these analyses clarify whether it is truly the integration of BEAST's insights with ML that drives BEAST-GB's success. Finally, we investigate whether BEAST-GB's powerful predictive abilities reflect mere flexibility in capturing idiosyncrasies in each dataset or a broader capacity to capture underlying choice tendencies. We do so by training it on data from some experiments and testing it on different experiments, thereby assessing its context generalization, a pinnacle of predictive modelling[20].

Results

CPC18, a choice prediction competition

Five of the present authors (O.P., R.A., E.E., M.T. and I.E., hereinafter the organizers) organized CPC18, a choice prediction competition for human choice between lotteries (https://cpc-18.com)[18], a domain that underlies both the foundations of rational economic theory[4,21] and the analyses of robust deviations from rational choice[2,3]. CPC18 used a unified space of decisions under risk, under ambiguity and from experience (Fig. 1), in which at least 14 classical choice anomalies emerge (including St. Petersburg's[1], Allais's[22] and Ellsberg's[23] paradoxes). Competing participants received choice data from 210 tasks sampled from this space and were required to predict the distribution of choices in 60 new held-out tasks sampled from the same space (without knowing which tasks would be used for testing during model development). Accuracy was measured by mean squared error (MSE), supplemented by a ‘completeness’ metric that is defined as the fraction of predictable variation in the data that the model captures[14]. It is calculated as $(\mathrm{MSE}_{\mathrm{random}} - \mathrm{MSE}_{\mathrm{model}})/(\mathrm{MSE}_{\mathrm{random}} - \mathrm{MSE}_{\mathrm{irreducible}})$, where $\mathrm{MSE}_{\mathrm{random}}$ is the test MSE of a model that assumes random behaviour, $\mathrm{MSE}_{\mathrm{model}}$ is the test MSE of the model and $\mathrm{MSE}_{\mathrm{irreducible}}$ is the estimated test MSE of a theoretical perfect model, whose error is only a result of the sampling variation (Methods). With the training data, the organizers also published (before the test data was collected) two baseline benchmarks: a purely behavioural model that is an adaptation of BEAST, and the hybrid model Psychological Forest[16] that uses the behavioural insights of BEAST as features in a random forest[24] algorithm (see the Supplementary Information for details).

Forty-six teams, involving 69 researchers representing 34 institutions from 16 countries, registered to the competition. A post-competition survey ($N = 29$; Supplementary Information) indicated that many teams invested considerable effort; the reported average time spent on model development was 66 h (s.d. 92). Twenty models were submitted in time. All submissions integrated behavioural assumptions, suggesting purely data-driven methods struggled in this domain.

The top-ranked submission, BEAST-GB, is an XGB algorithm[19] that uses the same features as the baseline hybrid Psychological Forest. BEAST-GB uses, in addition to features describing each task (hereinafter the ‘objective’ features), three sets of ‘behavioural’ features: (1) ‘naive’ features that capture naive intuition for what may matter in choice between lotteries (for example, the difference between the lotteries' EVs), (2) ‘psychological insight’ features that were hand-crafted on the basis of the behavioural insights underlying BEAST (for example, the difference between the probability of each lottery to generate a better outcome, based on BEAST's assumption of simultaneous mental sampling of outcomes from both lotteries) and (3) a ‘behavioural foresight’ feature: the numeric prediction of BEAST itself. Note the distinction between psychological insight features, designed to capture a general tendency that can drive behaviour, and behavioural foresight features, quantitative predictions of behaviour in a task

Table 1 | Features in BEAST-GB

Category and name	Descriptive label	Description
Objective		
Ha	High payoff A	H_A: high payoff in option A. When option A has multiple outcomes, Ha is the EV of the lottery in option A.
pHa	Probability high payoff A	pH_A: probability of Ha.
La	Low payoff A	L_A: low payoff in option A.
LotShapeA[a]	Shape lottery A	LotShape$_A$: shape of the distribution of the lottery in option A ('R-Skew', 'L-Skew' or 'Symm'). When option A does not have multiple outcomes, LotShapeA='–'.
LotNumA	Number of lottery outcomes A	LotNum$_A$: number of outcomes in distribution of the lottery in option A. When option A does not have multiple outcomes, LotNumA=1.
Hb	High payoff B	H_B: high payoff in option B. When option B has multiple outcomes, Hb is the EV of the lottery in option B.
pHb	Probability high payoff B	pH_B: probability of Hb.
Lb	Low payoff B	L_B: low payoff in option B.
LotShapeB	Shape lottery B	LotShape$_B$: shape of the distribution of the lottery in option B ('R-Skew', 'L-Skew' or 'Symm'). When option B does not have multiple outcomes, LotShapeB='–'.
LotNumB[a]	Number of outcomes lottery B	LotNum$_B$: number of outcomes in distribution of the lottery in option B. When option B does not have multiple outcomes, LotNumB=1.
Amb	Ambiguous task	Indicator for an ambiguous choice task (1 if true, 0 otherwise).
Corr[a]	Options correlation	Sign of correlation between generated payoffs in the two options (−1, 0 or 1).
block	Block number	The block number in repeated choice tasks (each block corresponds to 5 trials).
Feedback	Feedback block	Indicator for block with feedback (1 if true, 0 otherwise).
Dataset[a]	Experimental context	Dataset from which task is taken.
Naive		
diffEVs	Δ EVs	Difference between the payoff EV of option B and the payoff EV of option A.
diffSDs	Δ Standard deviations	Difference between the payoff standard deviation of option B and the payoff standard deviation of option A.
diffMins[b]	Δ Minimum payoffs	Difference between the minimal payoff of option B and the minimal payoff of option A.
diffMaxs	Δ Maximum payoffs	Difference between the maximal payoff of option B and the maximal payoff of option A.
Psychological		
diffBEV0	Δ Best EV estimates (no Fb)	Difference between the 'best estimate' of the EVs as per BEAST, before getting feedback. When the tasks are not ambiguous, diffBEV0 = diffEVs.
diffBEVfb	Δ Best EV estimates (with Fb)	Difference between the 'best estimate' of the EVs as per BEAST, after getting first feedback. When the tasks are not ambiguous, diffBEVfb = diffEVs.
pBbet_Unbiased1	Δ Probabilities better pay (no Fb)	Difference between the probability that option B provides better payoff than option A and the probability that option A provides better payoff than option B, as estimated by BEAST before getting feedback.
pBbet_ UnbiasedFB	Δ Probabilities better pay (with Fb)	Difference between the probability that option B provides better payoff than option A and the probability that option A provides better payoff than option B, as estimated by BEAST after getting feedback.
diffUV	Δ Uniform pay EVs	Difference between the EV of option B when all its outcomes are transformed to be equally likely and the EV of option A when all its outcomes are transformed to be equally likely.
pBbet_Uniform	Δ Probabilities better uniform pay	Difference between the probability that option B provides better payoff than option A and the probability that option A provides better payoff than option B, when both options are transformed so that their outcomes are equally likely.
RatioMin	Ratio minimum payoffs	Ratio between the smaller and the higher minimal outcomes of the two options. When the minimal outcomes have different signs, RatioMin=0.
SignMax[a]	Sign maximum payoff	The sign of the maximal possible payoff in the task (−1, 0 or 1).
diffSignEV	Δ Sign pay EVs	Difference between the EV of option B when all its outcomes are sign transformed and the EV of option A when all its outcomes are sign transformed.
pBbet_Sign1	Δ Probabilities better sign pay (no Fb)	Difference between the probability that option B provides better payoff than option A and the probability that option A provides better payoff than option B, as estimated by BEAST before getting feedback and after all payoffs are sign transformed.
pBbet_SignFB	Δ Probabilities better sign pay (with Fb)	Difference between the probability that option B provides better payoff than option A and the probability that option A provides better payoff than option B, as estimated by BEAST after getting feedback and after all payoffs are sign transformed.
Dom[a]	Dominant option	Trinary indicator for the option that stochastically dominates another ('1' = B dominates A; '−1' = A dominates B; '0' = neither option has dominance).
Foresight		
BEASTpred	BEAST prediction	The quantitative point prediction of BEAST for the choice task (and block). Predictions are made using the model's original implementation and without training it to new data (that is, using parameters as found in ref. 5).

This is an exhaustive list of every feature used in this Article as part of BEAST-GB. When run on particular datasets, some features may be completely constant and others may be duplicates of other existing features, in which cases these features are removed before running of the algorithm. [a]Categorical feature that is dummy coded before running of the algorithm. [b]diffMins belongs to both the naive and the psychological feature categories.

(cf. refs. 25,26). Table 1 details all features used. BEAST-GB achieved 92.6% completeness, capturing nearly all predictable variation in the test data and winning CPC18.

Analyses of feature importance. We investigated, using two methods, which features help BEAST-GB most in making such accurate predictions. First, we removed entire feature sets from BEAST-GB, retrained it and measured the drop in its predictive power. The results (Fig. 2a) show that removing the behavioural foresight feature, BEAST's prediction, led to the biggest decline, doubling the MSE and reducing completeness score to 82.8%. This highlights that BEAST alone provides accurate predictions of choice in CPC18. Removing the psychological insight features also degraded accuracy (MSE increased 18%, but completeness remained high). Recall these features were crafted on the basis of the assumptions of BEAST; the fact that removing them hurts performance despite the usage of BEAST itself as a feature implies that they hold information that extends beyond how they are captured in BEAST.

Second, we quantified the feature importances by computing their average absolute Shapley additive explanations (SHAP) values on the test set. SHAP, named after the Shapley value in cooperative game theory, is a popular way to compute feature importance in ML[27]. A feature's SHAP value captures its unique contribution to the model's prediction, such that larger absolute SHAP values imply greater importance. Figure 2b shows that the most important feature is the prediction of BEAST, followed by three psychological insight features. These insight features were designed to capture two assumed sensitivities: to the probability that one option provides a better payoff than the other, and to the difference between the lotteries' EVs or the EVs' 'best estimates' (BEAST also handles ambiguous lotteries in which the EVs cannot be computed; Supplementary Information). The results of both analyses thus suggest that the behaviourally informed features are vital for BEAST-GB's predictive power.

Which 'foresight' feature? Because the predictions of BEAST were the most useful feature in BEAST-GB, we next tested whether using the predictions of other behavioural models as foresight features would be as useful. We fitted four classical models of decisions under risk, including two variants of prospect theory, and used their predictions as features in an XGB trained to predict the competition's data. Notably, we did this for a subset of tasks that was the focus of classical decision research: pure decisions under risk (without ambiguity and without feedback). On this subset, BEAST, which derives its main assumptions from studies of the effects of feedback, should be at a considerable disadvantage. Unlike the other models, we also did not specifically fit it to this subset. Nevertheless, Extended Data Fig. 1 shows that BEAST is a far more useful 'behavioural foresight' than the other models. Using BEAST as foresight, completeness of the hybrid model is 90%. Replacing it with the next-best behavioural foresight, a version of cumulative prospect theory (CPT)[2], cuts completeness to 67%. Notably, the ML algorithm improved the accuracy of every behavioural model when its predictions were supplied as a foresight feature, yet even an XGB that combined all five foresights failed to outperform the single-BEAST configuration. Thus, the BEAST-derived foresight signal proved uniquely powerful.

Choices13k, behavioural theory when data are abundant

The results of CPC18 highlight BEAST's usefulness in predicting human choice between lotteries and demonstrate that integrating its predictions and behavioural insights into a ML algorithm yields further gains. In many real-world prediction problems, training data are rather limited, for example, because, when implementing a new incentivization policy, only few treatments can be piloted before choosing a policy. However, the training data in CPC18 are considerably smaller than in many tasks in which ML algorithms that are not ingrained with theoretical domain knowledge excel. This raises the question of whether behavioural theory remains necessary when the training data are large. It is possible that, with more data, purely data-driven methods can learn the regularities captured by BEAST (and/or other theories) directly, so behavioural insights matter only when data are scarce[17].

To explore this, we evaluated BEAST-GB on Choices13k, the largest publicly available dataset of human choice under risk and uncertainty[17]. It includes nearly 10,000 choice tasks similar to those used in CPC18 (see Methods and Table 2 for main differences). Importantly, using Choices13k, prior studies have shown that, with such large data, ML algorithms—specifically deep neural networks—could achieve high accuracy even without built-in behavioural logic (although training was more efficient with it)[7,17]. Following Peterson et al.[7], we repeatedly split the dataset into training (90%) and test (10%) sets and trained models on increasingly larger fractions of the training data. This procedure allows checking how much data are needed to reach different levels of predictive accuracy.

Figure 3 compares BEAST-GB with several benchmarks, including context-dependent (CD), one of the best and most expressive neural networks analysed in Peterson et al. BEAST-GB achieved state-of-the-art accuracy (MSE 0.00843), with 96.2% completeness, capturing nearly all predictable variation in the data. Furthermore, BEAST-GB required relatively few training examples to reach high accuracy: trained on just 2% of the training data (176 choice tasks), it already predicted more accurately (MSE 0.0110) than CD trained on all ~9,000 tasks (MSE 0.0113). This highlights how incorporating behavioural logic can dramatically improve sample efficiency, enabling models to achieve strong predictive performance with substantially less data.

Analyses of feature importance confirmed that behavioural features remain critical, even in this data-rich environment, although the influence of BEAST's predictions as a foresight feature diminishes with increasing data availability. Extended Data Fig. 2a shows that, when training data were scarce, removing BEAST's prediction feature sharply impaired performance, suggesting that BEAST provides an effective initial approximation that helps mitigate bias (see the Supplementary Information for bias-variance analyses). Yet, with sufficient data, the removal of the foresight feature was inconsequential (MSE 0.00853, not significantly different from BEAST-GB, $t(49) = -1.24$, $P = 0.22$, ΔMSE -0.0001, 95% confidence interval (CI) -0.0003 to 0.0001, Bayes factor favouring H_1 over H_0 (BF_{10}) 0.32). This suggests that, as training data increase, the model can learn a proper integration of the psychological insights underlying BEAST without direct access to BEAST itself. By contrast, removing psychological insight features—hand-crafted to reflect BEAST's behavioural mechanisms—reduced accuracy even with the full dataset (MSE 0.00879; significantly worse than BEAST-GB, $t(49) = -5.09$, $P < 0.001$, ΔMSE -0.0004, 95% CI -0.0005 to -0.0002). Furthermore, removing both the psychological insight and foresight features worsened performance still (MSE 0.00920), and using only objective task features drastically reduced accuracy (MSE 0.01530). Thus, even when data were abundant, purely data-driven models failed to fully capture the predictive power of behavioural insights. Analysis of SHAP values (Extended Data Fig. 2b) further supported these conclusions.

Interestingly, models trained using only behavioural features, without access to objective task structure, also performed significantly worse than BEAST-GB (MSE 0.00914; $t(49) = -8.08$, $P < 0.001$, ΔMSE -0.0007, 95% CI -0.0009 to -0.0005). This suggests that, while task structure alone carries little predictive power, it provides crucial context for behavioural features to be effectively leveraged. These results imply that some of BEAST-GB's success stems from an integration of task structure and BEAST's behavioural logic.

Using BEAST-GB to explain behaviour. If the predictive power of BEAST-GB involves successful integration of task structure and the behavioural insights of BEAST, it should be possible to identify classes of tasks in which BEAST-GB predicts systematically differently than BEAST. Because BEAST-GB captures nearly all of the predictable

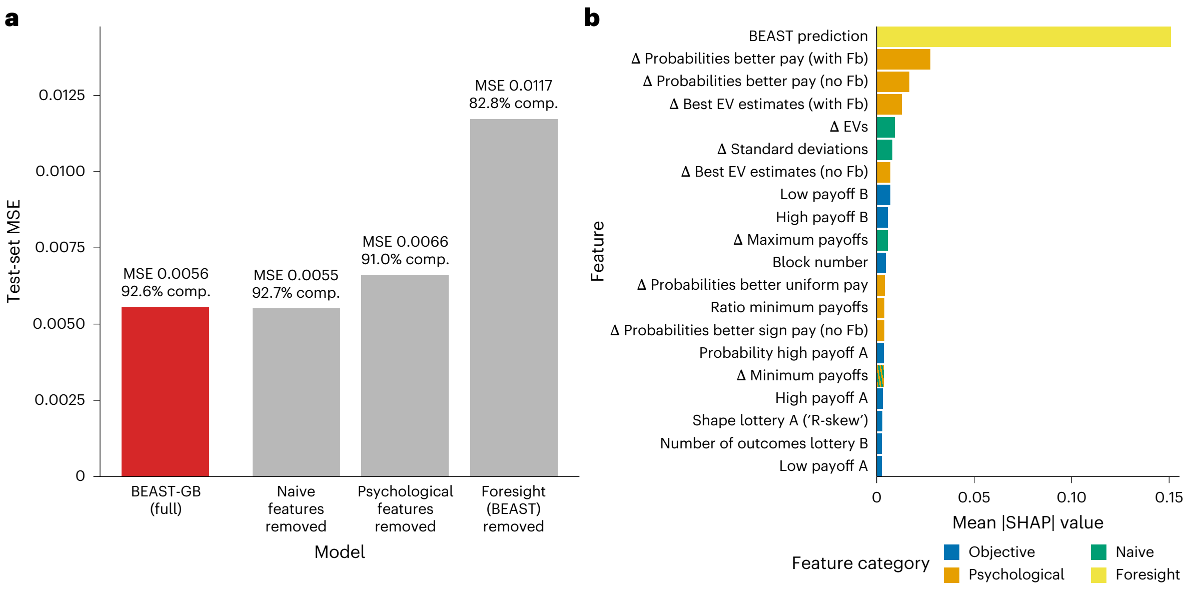


Fig. 2 | Feature importance analyses for CPC18 data. a, CPC18 predictive performance of BEAST-GB (in red) and variations of it that remove different feature sets. Bars show the MSE in the single held-out test-set ($n = 60$ tasks) per ablation. Completeness (comp.) (Methods), computed relative to a naive baseline (MSE 0.0637) and irreducible noise limit (MSE 0.0009). **b**, Average absolute SHAP values of BEAST-GB's features in predicting CPC18's test set, by feature category. For clarity, only the top 20 features are shown. Feature names and definitions appear in Table 1.

variation in the data, analysing where it deviates from BEAST can reveal choice patterns that BEAST overlooks, potentially informing improvements to the behavioural model itself. Note that this process is more effective than critiquing BEAST with respect to the data because its deviations from the observed behaviour also reflect unpredictable noise[15].

Our analysis of the differences between BEAST's and BEAST-GB's predictions showed that 90% of the variance in the deviations could be explained by three sets of intuitive corrections (Supplementary Information). First, BEAST's predictions are too extreme, especially when task complexity increases, implying BEAST-GB identifies that behaviour in the (online) experiments of Choices13k is noisier than BEAST (trained on laboratory data) expects[28]. Second, BEAST fails to capture systematic gain-seeking tendency in tasks that involve the possibility to avoid a sure loss. This behaviour contradicts loss aversion but is consistent with the experimental design in Choices13k, where negative payments were replaced with zero. Third, BEAST assumes that each of its mechanisms operates uniformly across all tasks, but BEAST-GB can dynamically adjust their relative influence based on task structure. Some of the systematic deviations of BEAST from BEAST-GB, such as gain-seeking in Choices13k, are probably dataset specific, but others can be more general. Indeed, insights from this analysis led to a simple correction to BEAST, which—without increasing complexity or reducing interpretability—improved its predictive performance across all datasets considered in this study (Supplementary Information).

These findings demonstrate how hybrid models like BEAST-GB not only enhance prediction accuracy but also serve as a powerful tool for refining and improving behavioural theories. By leveraging the flexibility of ML while preserving interpretability, BEAST-GB reveals systematic choice patterns that would otherwise be obscured by theoretical constraints or data noise.

HAB22, machine learning when theory fails

Although BEAST-GB achieved high accuracy in both CPC18 and Choices13k, these successes may have arisen primarily because BEAST itself is already very effective for those datasets. Indeed, in CPC18, the second-best submission was a minor modification of BEAST (Supplementary Table 1), and a scalable retrained variation of BEAST performed similarly to CD with completeness of 88.3% (ref. 29). Furthermore, the original BEAST, without retraining of its parameters, already captured much of the predictable variation in both CPC18 (88.9% completeness) and Choices13k (65.7%). Thus, it remains unclear how much real benefit comes from merging a strong behavioural model (BEAST) with ML, as opposed to simply relying on the behavioural model alone.

To investigate this question, we turned to a dataset of risky-choice tasks recently collected by He, Analytis and Bhatia, henceforth the HAB22 dataset[6]. HAB22 differs from CPC18 and Choices13k in several important ways (Table 2 and Extended Data Fig. 3). First, it includes data from multiple distinct experimental contexts. Second, it is restricted to choice between lotteries with up to two outcomes and without feedback. Last, it includes data from experiments designed to produce strong context effects[30]. In many of the tasks from these experiments, the lotteries' EVs differed dramatically, and—potentially because of context effects—participants often did not choose the option with the much higher EV[31,32]. While such tasks are useful for demonstrating interesting deviations from expected utility theory, using them can hurt BEAST's predictive power, as it assumes high sensitivity to EV differences. Thus, HAB22 allowed us to examine the generality of our results in several ways, as we show below.

Originally, HAB22 was used to evaluate 53 existing behavioural models by fitting them to each participant's data and then predicting the same individuals' choices on new tasks. BEAST-GB (like BEAST) was designed for predicting the behaviour of new decision-makers in new tasks, not for predicting that of known individuals. To evaluate how the framework presented here fares at the individual level, we developed two variants of BEAST-GB that retain its logic and rely on its population-level predictions, but use different learning algorithms and training regimes that better align with the small per-participant data involved. In the Supplementary Information, we demonstrate that these models predict the individual choices in HAB22 as well as or better than the best extant behavioural models. For consistency with our other analyses, however, here we compared BEAST-GB with the behavioural models in predicting aggregate choice rates for new participants facing new tasks.

Table 2 | Comparison between datasets used in this paper

	Dataset		
	CPC18	Choices13k	HAB22
Number of choice tasks	270	9,831	1,565[a]
Choice task properties:			
Number of trials in each task	25	5	1
Feedback after each choice?	First five trials without feedback, then full feedback	Full feedback	None
Number of outcomes in each lottery	Up to 10	Up to 10 in one lottery and up to 2 in the other	Up to 2
Ambiguity possible?	Yes	No	No
Number of tasks per participant	30	20	Varies between 46 and 150 (mostly consistent within experimental context)
Number of participants per choice task	At least 90	16 on average	Varies between 15 and 122 (mostly consistent within experimental context)
Location	Physical labs in the Technion and HUJI	Amazon Mechanical Turk	Physical labs in various locations (except Stewart15_1C_positive_skew, and Stewart15_1C_uniform, which was online)
Population	Mostly undergraduate students	MTurk workers	Students (Erev17app, Rieskamp_Positive, Stewart15_1A_negative_skew, Stewart15_1A_positive_skew, Stewart15_2A_negative_skew, Stewart15_2A_positive_skew, Stewart15_2B_negative_skew and Stewart15_2b_positive_skew, Stewart16) or pools of experimental participants (Fiedler12_exp1, Fiedler12_exp2, Pachur17, Pachur18_e1_session1, Pachur18_e1_session2, Stewart15_1C_positive_skew and Stewart15_1C_uniform)
Payment method	Payoff in 1 randomly selected task	Fixed proportion (10%) of payoff in 1 randomly selected task, but with minimal payoff of 0	Payoff in 1 randomly selected task (Erev17app, Fiedler12_exp1 and Fiedler12_exp2), fixed proportion of 1 randomly selected task (Rieskamp_Positive, Pachur17, Pachur18_e1_session1 and Pachur18_e1_session2), hypothetical (Stewart15_1C_positive_skew and Stewart15_1C_uniform) or contingent on performance but unclear from methods exactly how (Stewart15_1A_negative_skew, Stewart15_1A_positive_skew, Stewart15_2A_negative_skew, Stewart15_2A_positive_skew, Stewart15_2B_negative_skew and Stewart15_2b_positive_skew, Stewart16)

[a]When HAB22 is used for context generalization analyses, it includes 1,665 tasks. The 100 additional tasks come from an experimental context (Stewart15_1C_uniform) that includes many tasks that participants faced twice within a session and was removed for the analysis in which models were trained and predicted within contexts. Under context generalization, when models predict behaviour in new contexts, repeated choices were pooled together.

This comparison revealed that, on HAB22, the original BEAST (without retraining) fared poorly, achieving only 36% completeness. By contrast, the strongest purely behavioural model, a version of CPT, reached 93.8% completeness (MSE 0.0316). Nevertheless, as shown in Fig. 4, BEAST-GB, which used BEAST's very inaccurate predictions as a feature, outperformed all other models, with completeness of 94.8% (MSE 0.0307). This improvement over the best behavioural model was significant ($t(49) = 2.99$, $P = 0.004$, ΔMSE 0.0010, 95% CI 0.0003 to 0.0016). BEAST-GB's advantage persisted even when training and test sets used the same participants (predicting their aggregate choice in new tasks) and although the behavioural models (unlike BEAST-GB) were fitted to these participants' behaviour (Supplementary Information).

Interestingly, removal of the 'foresight' feature hurt the model's performance (MSE 0.0313, significantly worse than BEAST-GB, $t(49) = -4.08$, $P < 0.001$, ΔMSE −0.0006, 95% CI −0.0009 to −0.0003), and this feature remained the most important according to SHAP value analysis (Extended Data Fig. 4). Hence, BEAST holds important information concerning behaviour even when its raw predictions are poor. One reason for this is probably the high rank-order (Spearman) correlation ($\rho = 0.819$) between (the untrained) BEAST's predictions and the observed choice rates. In addition, we show in the Supplementary Information that most of the differences between the predictions of BEAST and BEAST-GB are accounted for when the mechanisms in BEAST are rescaled for each experimental context and task structure. This implies that the ML component in BEAST-GB identifies and corrects BEAST's context-dependent miscalibrations, enabling superior accuracy even when BEAST alone performs poorly.

Context generalization

The preceding analyses showed that, for each of three large datasets of human choice under risk and uncertainty, training BEAST-GB on tasks from within the same context yielded highly accurate predictions of new tasks. We next asked whether BEAST-GB, when trained on choice data from one experimental context, could effectively predict behaviour in a different experimental context. The ability to generalize across contexts—often called domain or context generalization—is a highly desirable property of predictive models[20,33,34]. Furthermore, recent work suggests that different experimental contexts in decisions under risk and uncertainty can systematically differ in subtle but important ways, meaning a model trained on one context can struggle when tested on another[28].

To examine BEAST-GB's capacity for context generalization, we exploited the fact that HAB22 is a collection of distinct experimental contexts. We systematically trained BEAST-GB on all but one of the contexts, then predicted behaviour in the held-out context, without using its choice tasks or participants during training. On average, BEAST-GB yielded MSE of 0.0162 in the unseen context, corresponding to 87.2% completeness (s.d. 0.08). That is, without access to data from the target context, BEAST-GB achieved over 87% of the predictive accuracy expected from a perfect hypothetical model that knows the population parameter for each task in that context.

Furthermore, 31% of the choice tasks in HAB22 appeared in more than one experimental context. This allowed us to compare BEAST-GB's generalization capacity (that is, its accuracy in predicting behaviour outside of context) with direct empirical generalizations, namely, with

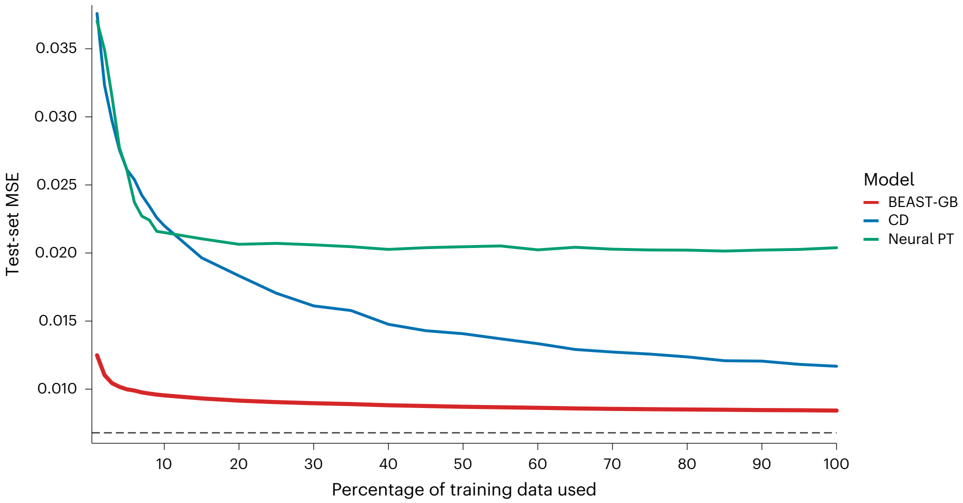


Fig. 3 | Test-set performance on Choices13k data. Data were split to 90% training (8848 tasks) and 10% held-out test data (983 tasks), and models were trained on fixed and increasing proportions of the training data. This process was repeated 50 times, and results reflect the average test-set MSE over these $n = 50$ train–test splits. The performance of Neural PT (neural prospect theory) and CD (an unrestricted deep neural network) is taken directly from the work of Peterson et al.[7]. The dashed line indicates the irreducible-error threshold (MSE 0.0068), which is the expected error of a perfect theoretical model (Methods).

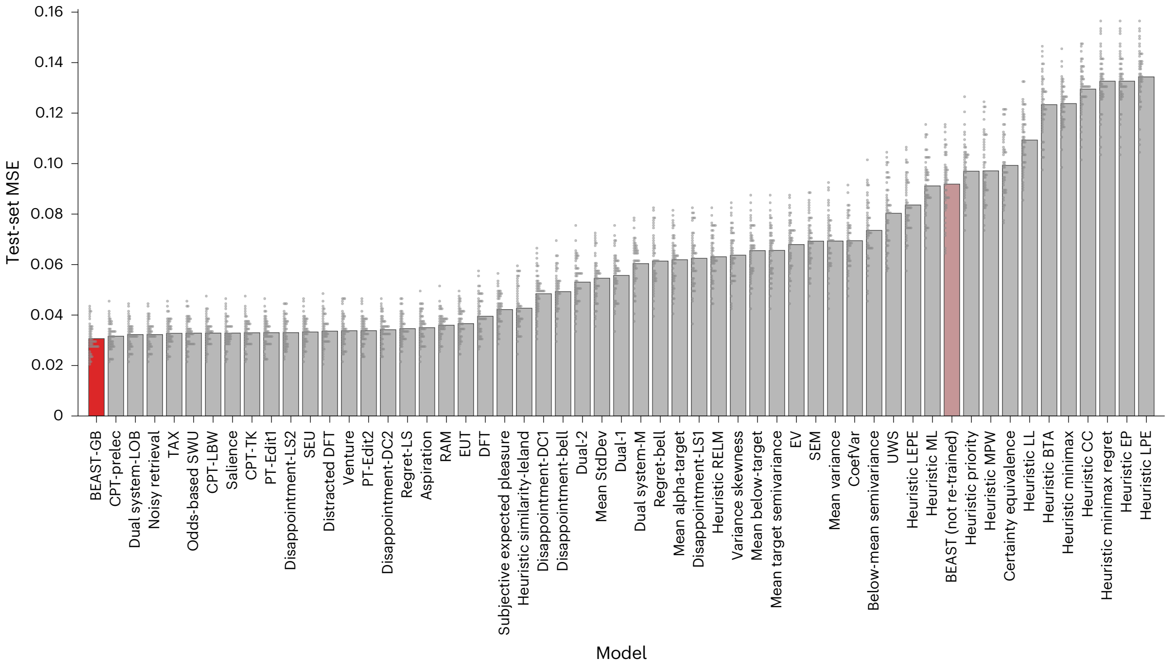


Fig. 4 | Test-set performance on HAB22 data. Performance is evaluated on the basis of tenfold cross-validation on choice tasks, and fivefold cross-validation on participants within experimental contexts. That is, models predict choice rates of new participants in new tasks (Methods). Bars show mean test-set MSE across the 50 nested cross validation folds ($n = 50$ fold-MSE values). Grey dots form a horizontal dot histogram of the fold-level MSEs (bin 0.001; stacked from left to right) for each model. Colours highlight BEAST-GB (red) and the original version of BEAST (rosy brown). For unabbreviated model names and model sources, see Supplementary Table 2.

using the observed choice rate of a given task in one context as a prediction to the choice rate of the same task in another context. Note that, because people in different contexts do not necessarily behave similarly and given sampling errors, quantitative models trained to capture general patterns of behaviour across tasks might predict more accurately the choice rate in the new context. That is, the error of the predictive models could potentially be smaller than the average sampling error. As Fig. 5 shows, none of the behavioural models examined by He et al. achieved this feat, but BEAST-GB did. Its MSE when predicting choice rates of known tasks in new experimental contexts was 0.0121 (91.8% completeness), representing a 13% improvement over simply assuming behaviour directly generalizes across experimental contexts and predicting the same task's observed choice rate from the training contexts. The difference is significant, $t(827) = -3.79$, $P < 0.001$, ΔMSE −0.0019, 95% CI −0.0028 to −0.0009. That BEAST-GB usefully predicts choice behaviour in new contexts it was not trained on suggests that it captures generalizable choice tendencies, rather than merely fitting idiosyncratic patterns from specific samples of tasks and participants[34].

Discussion

Our Article introduces BEAST-GB, a hybrid model that integrates a strong behavioural theory (BEAST) with ML to predict human choice

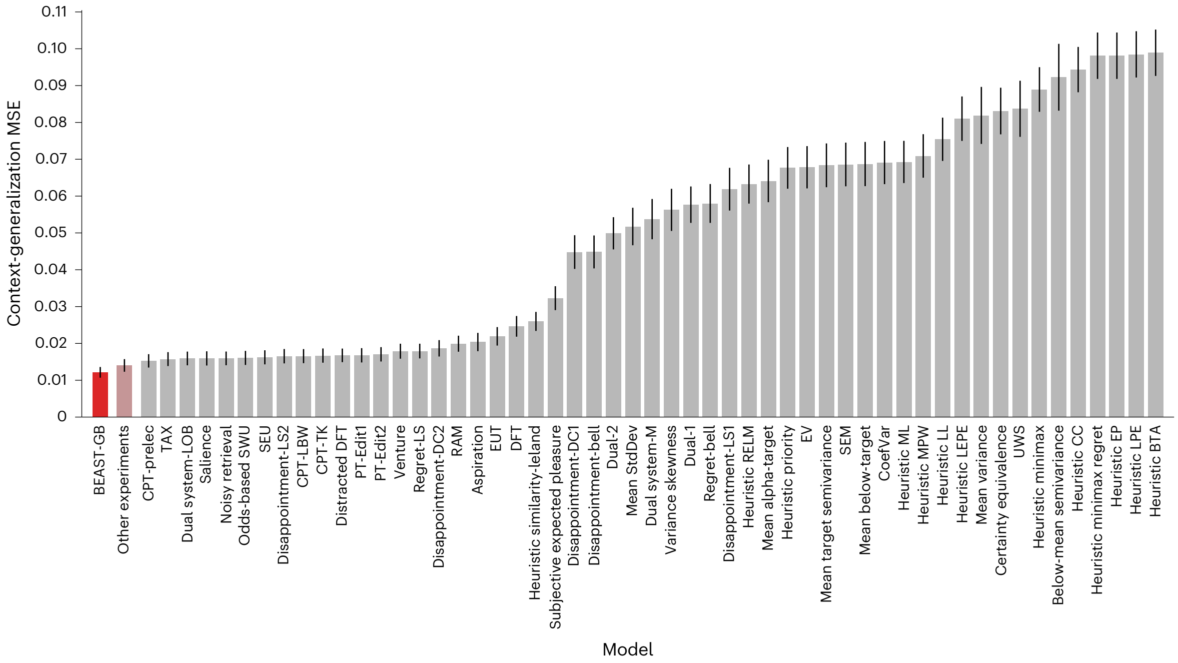


Fig. 5 | Predictive accuracy in context generalization task. Training data always include 15 experimental contexts to predict the behaviour in a 16th context in HAB22 data. Predictions are made for the 828 instances where a choice task that appears in a test experimental context also appeared in one or more of the train contexts. 'Behavioural models' (grey) prediction is the average training prediction (that is, best fit) in the target task across all participants in the training data. 'Other experiments' (rosy-brown) prediction is the average observed behaviour across all participants in the training data of the target task. BEAST-GB is highlighted in red. Bars indicate MSEs across the $n = 828$ tasks. Error bars are 95% CI for the mean. See Supplementary Fig. 1 for full error distributions. See Supplementary Table 2 for unabbreviated model names and model sources.

under risk and uncertainty. Across three datasets encompassing more than 11,000 choice tasks, BEAST-GB demonstrated state-of-the-art predictive accuracy, consistently capturing over 92% of the predictable variation that would have been captured by a perfect (hypothetical) model. BEAST-GB won an open prediction competition featuring genuinely independent test data[35,36] and maintained predictive superiority within the largest public dataset of risky choice as well as a collection of 15 distinct experimental contexts that differ in participant pools, settings and methodologies. Furthermore, BEAST-GB successfully generalized across contexts, outperforming even direct empirical generalizations from observed behaviour. These findings underscore BEAST-GB's broad applicability across decision-making environments.

Our analyses suggest that BEAST-GB's predictive success stems from the effective synergy between behavioural theory and ML. The integration involves three categories of features: objective task characteristics, psychological insights that represent BEAST's behavioural mechanisms, and foresight provided by BEAST's quantitative predictions. The foresight feature supplies the ML algorithm with an initial powerful signal about likely behavioural patterns. The ML component then adjusts these predictions by dynamically weighting the psychological insights underlying BEAST across diverse task structures. Fully disentangling these adjustments is difficult, so using our hybrid approach purely to explain the cognitive processes behind choice behaviour remains limited. Nevertheless, we show that some of the adjustments can be explicitly identified, offering insights into systematic limitations within BEAST itself. Notably, these insights led to a refinement of the original behavioural model, highlighting that such hybrid models can also serve as theoretical diagnostic tools capable of improving our understanding of decision-making processes.

A notable advantage of the hybrid approach is that it provides an efficient way to scale rigid and complex behavioural models like BEAST to new and large datasets without extensive refitting. BEAST, by itself, involves computationally demanding simulations that make training on new data cumbersome. BEAST-GB circumvents these limitations, rapidly adjusting BEAST's theoretical predictions and insights to new contexts. Furthermore, because BEAST was developed to capture behaviour across a wide set of situations, it includes rigid constraints that are not necessarily theoretically grounded but help it avoid overfitting. Some of these constraints, however, may limit BEAST's adaptability. For example, BEAST gives high weight to the best estimates of the EVs, contributing to its lower predictive accuracy in some contexts (specifically some experimental contexts in HAB22 that showcase low sensitivity to EVs)[37]. BEAST-GB utilizes the information embedded in BEAST while effectively avoiding this bias. This demonstrates that the scalability afforded by hybrid models can advance behavioural research by enabling exploration of complex phenomena without the constraints imposed by the behavioural models' architectural rigidity.

Another strength of our hybrid approach is its generalizability. Indeed, the approach underlying BEAST-GB is not restricted to predicting choices between lotteries. In the Supplementary Information, we demonstrate that a similar hybrid approach can achieve state-of-the-art predictive accuracy in an entirely different decision domain, two-player extensive form games. For this approach to be effective in other domains, the key requirement is a foundational behavioural model that is reasonably accurate and sufficiently broad to provide meaningful behavioural insights that ML can exploit and extend. Hence, our hybrid approach is naturally limited by the theoretical basis and generalization capacity of extant models in each behavioural domain.

Implications for research in behavioural science

Behavioural science predominantly seeks explanations of behaviour, leading many researchers to focus on discovering new phenomena and specifying causal mechanisms[38]. Although such work is invaluable, it inevitably leads to the study of narrow scenarios designed to illustrate specific phenomena. For example, behavioural decision-making research focuses on situations that demonstrate deviations from rational choice.

Moving from elegant but narrow explanations towards robust and useful predictions—essential also for validating underlying theoretical mechanisms— requires greater emphasis on identifying behavioural principles that reliably generalize across broader sets of tasks.

We speculate that this rationale explains why BEAST, and its underlying mechanisms, provide highly useful behavioural insights for BEAST-GB. Unlike classical models primarily designed to capture anomalies where the rational benchmark is obvious, BEAST was originally developed to predict choice across a broad set of situations, including decisions under ambiguity and from experience. Its main assumptions are grounded in fundamental learning processes (for example, that choice is sensitive to the probability of obtaining better outcomes), many of which are shared across species[39,40], highlighting their potential generality and robustness. By considering a broad spectrum of situations, BEAST's developers could identify generalizable and useful insights that proved critical in enhancing BEAST-GB's predictive robustness and applicability.

Conclusion

Our research advances behavioural decision-making research by demonstrating the power of hybrid models that integrate behavioural logic with ML. BEAST-GB's success across diverse datasets and tasks and its ability to generalize across experimental contexts sets a new benchmark for accuracy and generalizability in the field. Looking forward, the integration of theoretical insights derived from a prediction-focused approach to behavioural science[41] with ML offers a promising avenue for developing more adaptable, accurate and generalizable models of human behaviour.

Methods

Model evaluation

Throughout the Article, we evaluated models using their MSE between the predicted and the observed choice rates across tasks in the (test) data. The MSE was recently recommended as the preferred measure for evaluation of behavioural models as it satisfies all desired properties of a loss function in this domain[42]. In addition, to help interpret the accuracy of the models, and following the suggestion of Fudenberg et al.[14], we computed the models' completeness score, measured as the proportion of predictable variation in the data that the model captures. Completeness equals $(\text{MSE}_{\text{random}} - \text{MSE}_{\text{model}})/(\text{MSE}_{\text{random}} - \text{MSE}_{\text{irreducible}})$, where $\text{MSE}_{\text{random}}$ is the MSE of random guessing (as defined by Fudenberg et al.[14]), $\text{MSE}_{\text{model}}$ is the MSE of the model in question, and $\text{MSE}_{\text{irreducible}}$ is an irreducible error, that is, the portion of the total error considered unpredictable. To get $\text{MSE}_{\text{irreducible}}$, we aimed to estimate the expected MSE of a perfect hypothetical model that accurately predicts the population choice rate in a task. Notably, the computed MSE of such perfect theoretical model would probably be positive because models are evaluated based on their accuracy in predicting estimates of the population choice rates, namely the observed sample choice rates. That is, the observed error of a perfect theoretical model in task i is the sampling error, and thus the computed MSE of this model is equal to the average (over choice tasks) of the squared sampling errors. As the expectation of the squared sampling error equals the variance of the sample average, we get

$$\text{MSE}_{\text{irreducible}} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\mu_i} - \overline{x_i}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\mu_i - \overline{x_i}\right)^2,$$

$$E(\text{MSE}_{\text{irreducible}}) = \frac{1}{N}\sum_{i=1}^{N}E_i\left(\mu_i - \overline{x_i}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\text{Var}\left(\overline{x_i}\right) \cong \frac{1}{N}\sum_{i=1}^{N}\frac{S_i^2}{n_i}$$

where $\widehat{\mu_i}$ is the prediction of the perfect hypothetical model for task i, $\overline{x_i}$ is the observed choice rate in the sample for task i, μ_i is the true population choice rate, S_i^2 is the sample variance of task i, n_i is the sample size for task i, and N is the number of choice tasks. That is, we estimated $\text{MSE}_{\text{irreducible}}$ as the average of the squared standard errors.

BEAST-GB model

BEAST-GB is an XGB algorithm that uses the features detailed in Table 1. Most features used by BEAST-GB are derived from the behavioural model BEAST designed to predict human decision-making under risk and uncertainty at the population level[5].

Theoretically, BEAST is grounded in the idea that people adapt strategies that proved effective in past situations perceived as similar to the current one[43–45]. It assumes that individuals act as intuitive classifiers: a current task is classified alongside similar previous ones, and a strategy that worked well in that class is invoked[46]. Because the classification can be imperfect, the chosen strategy may sometimes be ill-suited to the current context, resulting in behavioural 'biases'. Instead of explicitly modelling this complex, individual and idiosyncratic classification process, BEAST approximates its main implications for the aggregate behaviour in risky and uncertain decisions by assuming people in these contexts primarily rely on five cognitive strategies. These strategies are choosing options that (1) are best in expectation, (2) minimize immediate regret, (3) maximize the chances to get a better payoff sign, (4) maximize the worst possible payoff and/or (5) yield a better payoff if all outcomes were equally likely. The output of the first strategy is computed explicitly or on the basis of one's 'best estimate' of the EV (if direct computation is impossible). The output of the other four strategies is implemented via a mental sampling process involving potentially biased 'sampling tools' (see the Supplementary Information for the implementation details). Each of the five cognitive strategies was previously translated into psychological insight features[16], which are now used in BEAST-GB.

XGB[19] is an algorithm that efficiently and effectively implements the idea of gradient boosting. Gradient boosting is an iterative ensemble procedure in which simple regression trees—models that repeatedly split the data on the basis of threshold conditions, thereby creating piecewise-constant predictions—are added one at a time to reduce the errors of the existing model. Each new tree learns to predict the residuals from the previous round, so that, over many iterations, the ensemble flexibly models nonlinearities and interactions among features, in a context-dependent manner. To reduce overfitting and improve generalization, XGB includes additional regularization, as well as random selection of features to be used in each iteration. In BEAST-GB, the algorithm takes as input both the objective features that capture the structure of the task and behavioural features that capture behaviourally relevant properties of the task. The algorithm then iteratively learns when and how to utilize them, searching at each iteration for feature interactions that best reduce the remaining prediction errors. The result is an ensemble that effectively ties together the signals available in the various features.

We implemented the following pipeline to train BEAST-GB on each choice dataset. First, we generated the features for each choice task. This notably includes generating the choice rate prediction of the original BEAST model for that choice task. Note that BEAST is not refitted to the new data. Its predictions (to be used as foresight feature) are derived using the original values of parameters fitted to the training set of CPC15 (Supplementary Information)[5]. Second, we coded categorical features to numeric using dummy coding. Third, because in particular datasets some features may turn out completely constant and/or duplicates of other features, we removed such features from the data. Fourth, we randomly split the data to a train and a held-out test set (unless the data were already organically split, like in CPC18). Fifth, we standardized all features by subtracting their average and dividing by their standard deviation in the train set. Sixth, we tuned the algorithm's hyperparameters using five repetitions of fivefold cross-validation implemented on the train set (see Supplementary Table 3 for the values of the hyperparameters in each dataset). Finally, we trained the algorithm on the full train set with the chosen hyperparameters and generated its predictions for the held-out test set.

Feature importance analyses

Throughout the Article, we assessed the relative importance of features included in BEAST-GB for prediction using two distinct methods. The first involved systematically removing sets of features from the tuned model, retraining it on the train set and evaluating the predictions of the new model (that is, without the removed features) on the test set. The second method involved computing the mean absolute SHAP values (using package SHAPforxgboost[47] in R) over all predictions of the test set (or, when models were evaluated using multiple iterations using different test sets, all predictions of the test sets).

CPC18

Experimental task. Similar to the paradigm used in CPC15[5], the experimental paradigm in CPC18 involved binary choice under risk, under ambiguity and from experience. As seen in Fig. 1, decision-makers were presented with two lotteries (option A and option B) and were asked to choose between them repeatedly for 25 trials. In the first five trials, they did not get any feedback, but starting from the sixth trial, they received full feedback concerning the outcomes generated by each option (both the obtained and the forgone payoffs were revealed). Choice options in CPC18 may include up to ten outcomes, may involve ambiguity (that is, probabilities of potential outcomes of one of the options were not revealed to the decision-maker) and may be correlated between them. A choice task is thus uniquely defined by 12 dimensions: 5 determine the outcome distribution of option A (L_A, H_A, pH_A, LotNum$_A$ and LotShape$_A$), five determine the outcome distribution of option B (L_B, H_B, pH_B, LotNum$_B$ and LotShape$_B$), one (Amb) determines if the task involves ambiguity and one (Corr) determines whether the outcomes in the two options are correlated. See the Supplementary Information for more details on these dimensions and how they define the tasks.

The space of choice tasks that is implied by these dimensions extends the space studied in CPC15 by allowing both options (rather than just one) to have up to ten outcomes. Within this space, it is possible to replicate 14 classical behavioural decision-making phenomena[5]: the Allais' paradox[22], the reflection effect[3], overweighting of rare events[3], loss aversion[3], St. Petersburg's paradox[1], Ellsberg's paradox[23], the elimination of loss aversion at low magnitudes[48], the break-even effect[49], the get-something effect[50], the splitting effect[51], underweighting of rare events[52], the reversed reflection effect[52], the payoff variability effect[53] and the correlation effect[54].

Experimental data. The data used in CPC18 includes 694,500 decisions made by 926 different decision-makers across 270 binary choice tasks. Tasks were divided into nine cohorts. Each decision-maker faced one cohort of 30 tasks in random order and made 25 choices in each task. The first five cohorts were also used in CPC15[5], and details on these data are provided elsewhere. The choice tasks in the four additional cohorts were randomly selected from the space of tasks investigated in CPC18 according to a predefined task selection algorithm (Supplementary Information). Two cohorts of choice tasks were then run in each of two new experiments that used the same participant pool and a very similar design to those used for CPC15.

Each experiment involved 240 participants (experiment 1: 139 females, $M_{\mathrm{Age}} = 24.5$, range$_{\mathrm{Age}}$ 18–37; experiment 2: 141 females, $Mean_{\mathrm{Age}} = 24.7$, range$_{\mathrm{Age}}$ 18–50), mostly undergraduate students, participating in one of two (physical) lab locations: the Technion and the Hebrew University of Jerusalem. No statistical methods were used to predetermine sample sizes, but our sample sizes are larger than those used in previous publications focusing on predictions of choice under risk and uncertainty[5,6,17]. Informed consent was elicited from all participants at the beginning of the experimental session. The experiment lasted approximately 45 min. Participants were paid for one randomly selected choice they made, in addition to a show-up fee. The final payment ranged from 10 to 136 shekels, with a mean of 40 (about US$11) for experiment 1 and from 10 to 183 shekels, with a mean of 41.9 for experiment 2. The experiments complied with all ethical regulations and were approved by the Social and Behavioral Sciences Institutional Review Board in the Technion and by the Ethics Committee for Human Studies at the Faculty of Agriculture, Food, and Environment at the Hebrew University of Jerusalem.

Competition procedures and protocol. In May–June 2017, the organizers ran experiment 1. They then used the combined data from experiment 1 and from CPC15 to develop their baseline models (Supplementary Information) and made the data publicly available. In January 2018, they published a call to participate in the competition in major mailing lists and on social media. The competition included two independent challenges, and in this Article, we focus on the first (see the Supplementary Information for details on the second). In that challenge (or track), the goal was to provide, for each of 60 choice tasks from experiment 2, run in June–July 2018, a prediction for the progression over time of the mean aggregate choice rate of one of the options. Specifically, the 25 trials of each task were pooled to five blocks of five trials each, and the goal was to predict the mean aggregate choice rates of option B in each of the five time blocks. As the exact nature of the tasks was unknown to modellers at the time of model development, a competing model was required to get as input the values of the 12 dimensions defining each task and provide as output a sequence of 5 predictions (each in the range 0–1) for the mean choice rates in that task.

Interested participants were required to register for the competition in advance. Each person could register as a (co-)author of no more than two submissions per track and be the first author of no more than one submission per track. In addition, each person could make one additional early-bird submission, sent to the organizers by the end of January 2018. Submissions had to be made on or before the submission deadline (24 July 2018). In practice, this meant sending the organizers a complete, functional, documented code of the submission. The code could have been written in Python, R, MATLAB or SAS. The code was required to read the dimensions of a choice task and provide as output a prediction for the choice rates in the five blocks. One day after the submission deadline, the organizers published the test-set tasks (the 60 tasks from experiment 2). That is, submissions were blind to the tasks on which they were tested. Participants then ran their code on the test-set tasks and submitted the predictions. Finally, the organizers published the data to be predicted so participants could evaluate their prediction error. The organizers verified that the code for each of the top ten submissions produces the reported predictions and published the results.

Statistical significance. Because ranking of submitted models may depend on the (random) selection of the competition's test-set tasks, we used a bootstrap analysis (using Package boot[55] in R) to compare each submitted model with the competition's winner BEAST-GB. Specifically, we simulated 10,000 sets of 60 test choice tasks each by sampling with replacement from the original test set, computed the MSE of each submission in each simulated set and then counted the number of sets in which a submitted model outperformed BEAST-GB. The proportion of test sets in which a model would have outperformed the winner is the estimated P value for the difference between the winner and the model: If it is smaller than 0.05, then BEAST-GB is considered to predict significantly better.

Foresight comparison analysis. To compare the value of using BEAST as a foresight feature with the value of using other classical decision models as foresight features, we used a subset of the CPC18 data that includes only decisions under risk without feedback: choices made in tasks without ambiguity in the first block of five trials in each choice task. There were 230 such tasks. Each model, except BEAST, was fitted to the aggregate choice rates of the 182 of these tasks that were part of CPC18's training data, using a grid search over the parameter space.

BEAST was not fitted to these data. The values of its free parameters reflect the best fit to all five blocks of all 90 training problems from CPC15[5], which are a subset of CPC18 training data. The models then all predicted the aggregate choice in the 48 remaining tasks that were part of CPC18's competition data. Finally, we used those predictions as a foresight feature in XGB algorithm with hyperparameters tuned according to CPC18's train set subset of decisions under risk tasks without feedback. As additional features (beyond the foresight feature), we used the set of objective features that define each choice task. In this exercise, BEAST was compared with two versions of CPT[2], with the priority heuristic[56] and with the decision-by-sampling model[57]. In addition, we also compared it with an 'ensemble' model that includes all five foresight features (that is, the predictions of all five behavioural models were used as features in addition to objective features). The Supplementary Information provides details on the implementation of the various models and detailed results.

Choices13k

Data. The Choices13k dataset was originally presented by Bourgin et al.[17] and includes 13,006 binary choice tasks. Tasks were generated by the task generation algorithm used in CPC15[5] and are therefore all members of the same space used in CPC18 that extends it. Hence, they can all be described by the set of objective features in Table 1. Specifically, each choice task includes two options marked A and B, between which participants in an online experiment chose repeatedly across five trials. The data include, for each choice task, the proportion of times in which participants chose option B.

Participants in the experiment, Amazon Mechanical Turk users, were each presented with 20 choice tasks. On average, each task was faced by 16 participants. Participants were paid US$0.75 plus a 10% bonus on their winnings from one randomly selected task, unless their payoff was negative, in which case the bonus was set to zero. As in the work of Peterson et al.[7], we removed from the dataset tasks in which one of the options was ambiguous and tasks in which participants did not receive any feedback, resulting in a dataset containing 9,831 risky choice tasks in which participants made five consecutive choices with full feedback after each choice. Additional details of this dataset can be found in the works of Peterson et al.[7] and Bourgin et al.[17]. Extended Data Fig. 3 provides a visual representation of the wide coverage of this dataset, particularly in comparison with the data of CPC18. Table 2 summarizes the main differences between Choices13k and CPC18.

Benchmark models. We compared BEAST-GB with models developed in the work of Peterson et al.[7] that includes details of these models. In particular, we present the performance of BEAST-GB in comparison with the performance of two models from that study: Neural PT and CD. Neural PT is a neural network stochastic variant of prospect theory[3] in which the model searches the entire class of possible payoff and probability transformation functions assumed in prospect theory. Note that the search is over not only the space of parameters of the functions but also the functional forms themselves. In a sense, Neural PT reflects the version of prospect theory that best captures the data, and Peterson et al. show that it indeed predicts better than many other variations of prospect theory (including CPT). CD is the model that (after sufficient training) performed best in Peterson et al.'s analysis of this data. It is a fully unconstrained neural network that takes all information about both gambles as input and produces the choice rate as output. Because it is unconstrained, it effectively allows the network to learn subjective transformations of both outcomes and probabilities of the gambles, but in ways that are sensitive to the context of the other gamble. The performance of these benchmark models was taken directly from the analysis in Peterson et al.[7].

Error evaluation. To evaluate the models' error in Choices13k, we followed the original pipeline used by Peterson et al.[7]. Specifically, we performed 50 iterations of the following process. First, we split the data to 90% train set and 10% test set (choosing 983 choice tasks randomly for the latter). Then, we trained the model on an increasing proportion of the train set, ranging between 1% of the train set (88 choice tasks) to 100% of it (8,848 choice tasks). Next, we used the trained model to predict the held-out test set and computed its MSE. That is, for each proportion of the train set, we computed 50 MSEs on the test set. The reported results are the average of these 50 MSEs. To statistically compare the performance of different models, we used paired *t*-tests over the resulting MSEs (for 100% of the training data).

To derive the predictions of the model for further analysis, we performed five repetitions of a tenfold cross-validation procedure, so that each task's prediction was based on the average of exactly five predictions of BEAST-GB, each derived when the algorithm is trained on (a different) set of 90% of the data.

Using BEAST-GB to explain behaviour. We probed the differences between the predictions of BEAST and those of BEAST-GB in an iterative process of scientific regret minimization[15], a process in which the theoretical model is critiqued with respect to a more predictive but less interpretable model. The idea underlying this process is that errors of the theoretical model can result both from it missing predictable patterns and from noise. Because BEAST-GB predicts almost all predictable variation, using it to critique BEAST is more effective than using the (noisy) data itself, especially because BEAST-GB is a derivation of BEAST.

In each iteration, we sorted the tasks by descending order of the squared error between the two models' predictions. We then examined the tasks with the largest errors, trying to identify what features of behaviour BEAST-GB captures, but BEAST does not. Upon identifying a pattern, we linearly corrected the predictions of BEAST so that they were closer to those of BEAST-GB and then moved to the next iteration. To avoid increasing BEAST's complexity and reducing its interpretability, most of these corrections were statistical: we changed only the predictions of BEAST after they were derived. However, we also found a possible mechanistic correction to BEAST (changing the model itself before deriving its new predictions) that does little to the model's complexity and interpretability. We then implemented this correction, trained the new version of the model on the CPC18 training data and derived the trained model's predictions for all three datasets we use in this Article (Supplementary Information).

HAB22

Data. HAB22 includes data assembled by He et al.[6] from 15 different experimental contexts. The data from these different contexts were originally published in seven distinct papers by various researchers[5,30,58–62]. In each experimental context, participants made multiple one-shot choices between binary lotteries with up to two outcomes without feedback. Hence, the experimental task here was different from that used in CPC18 and Choices13k. Moreover, some choice tasks in this dataset are very different from the tasks in the other two datasets. Specifically, the difference between the EVs of the lotteries in some choice tasks here is especially large. For example, one task involved a choice between 500 with probability 0.4 versus 50 with probability 0.8 (EV difference of 160), and another task involved a choice between 500 with probability 0.8 and 100 for certain (EV difference of 300). In both tasks, most participants failed to maximize EV. Extended Data Fig. 3 shows a two-dimensional visualization of the similarities and differences between all choice tasks used in this Article and highlights that in HAB22 there is a cluster of choice tasks very different from the rest. Table 2 presents further details on this dataset and compares its main properties with those of the other datasets. In total, the HAB22 data include 1,565 choice tasks, although some of these are identical but were run in different experimental contexts and are thus treated as distinct.

Originally, He et al.[6] used four additional experimental contexts in their analyses. However, the data in these contexts are not usable for the purpose of our model comparisons[37]. In three contexts, there was an indexing error resulting in mismatches between the task IDs in the raw data and the original task IDs. This unfortunately has led to a mismatch between the parameters defining each task and the choice rate associated with it in the data. Consequently, the measured performance of the behavioural models that He et al. trained was distorted. In a fourth context, participants faced many of the same choice tasks more than once. As a result, the same exact task was often included both in the train and test set of the behavioural models. Hence, we could not properly compare BEAST-GB with the behavioural models in these four contexts and chose to exclude them.

Benchmark models. We compare BEAST-GB to all 53 behavioural models that He et al. investigated for the mixed gambles domain (Supplementary Table 2). Models are diverse and include a range of different assumptions about human risky choice. Details of these models can be found in the work of He et al.[6]. Under He et al.'s inclusion criteria, all behavioural models had to include precise functional forms that have analytically specified likelihood functions. This allowed fitting of each model to each individual in each experimental context separately. Yet, this also excluded the model BEAST whose prediction is used as a feature in BEAST-GB. Hence, we also derived the predictions of BEAST, without retraining of its parameters, and present them for comparison. As an additional benchmark, we also trained behavioural-theory-free deep neural networks and report on them in the Supplementary Information.

Evaluation method. In their original investigation, He et al. fitted each of the behavioural models to each individual participant separately, using a subset of the choice tasks that the participant faced, and evaluated the fitted models on the basis of their ability to predict the choices of the same individuals in the other (test) choice tasks. We consider this 'known' individuals prediction task in the Supplementary Information. BEAST-GB is a model for the prediction of new (unfamiliar) participants in new choice tasks (and the best prediction for new participants is the prediction of the mean choice behaviour of the population). Thus, and to be consistent with the rest of the current study, we evaluated the models on the basis of their ability to predict the choice rates of a new sample of participants from the population (that is, participants that the model had no access to during training) in new choice tasks. Hence, we first split the participants in each of the 15 experimental contexts to five folds. We then repeatedly used data of four folds of participants for training and predicted the data of participants in the last fold. This was done in addition to using He et al.'s original segmentation of the choice tasks in that experimental context to ten folds, using only choice tasks in nine of these folds for training and predicting behaviour in the tenth fold. That is, the train data included choices of 80% of the participants in 90% of the choice tasks, whereas the test data included choices of the other 20% of participants in the other 10% of the choice tasks of each experimental context.

As He et al. derived individual participant predictions for each choice task in each benchmark behavioural model, we averaged these original individual predictions across the participants in the train set to derive a prediction for the aggregate out-of-sample choice rate in the test-set task. BEAST-GB was trained on the aggregated choice rates in the training data (that is, unlike the benchmark behavioural models, BEAST-GB did not use individual participant data for its training). This process was repeated 50 times with different combinations of participants and tasks for the test set (that is, we essentially performed a double cross-validation procedure, on participants and on tasks). The reported results are the average of these 50 runs. To statistically compare the performance of different models, we used paired t-tests over the 50 resulting MSEs.

Context generalization

In the analyses of context generalization, we used the HAB22 data, with the addition of another experimental context ('Stewart15_1C_uniform') that we previously excluded because in that experiment many tasks were faced by the same participants more than once. Thus, when models were trained and tested within context, using this additional context introduces data leakage: The train and test data include choices of the same people in the same tasks. Under context generalization, however, models always predicted out of context and so there were no data leakage issues. Hence, here, we used 16 experimental contexts. Excluding this dataset does not qualitatively change any of the results.

Specifically, we repeatedly trained BEAST-GB on exactly 15 experimental contexts and then generated its predictions for the 16th context. Note that the model could not use the dataset feature here, as its values differed between training and testing. We report on the model's performance in this task of context generalization in two ways. First, we simply computed the MSE and completeness of the model in each of the 16 unseen datasets separately and report the average of these 16 MSEs and completeness scores.

The second evaluation we used relies on the fact that the exact same choice tasks (that is, choice between the same two payoff distributions) were at times used in different experimental contexts in HAB22. Specifically, there are 1,221 unique choice tasks in HAB22, and 384 of these were independently used in more than one experimental context: 338 tasks were used in two contexts, 33 were used in three contexts, 12 were used in four contexts and 1 task was used in five contexts. Thus, there were 828 instances where a choice task from the test set (the 16th experimental context) also appeared in the train set (at least once). For each of these 828 instances, we computed the prediction errors of BEAST-GB, and we report on the MSE across all these instances.

In addition, we computed the prediction error of an non-parametric model that predicts, in each instance, the observed choice rate of the same task in the training data. This allowed us to evaluate the error of BEAST-GB relative to a very strong benchmark that assumes behaviour in the same task is similar across experimental contexts. Note that the expected error of this benchmark is the sampling variance, and so a model whose average prediction error is smaller than the average sampling variance should be more accurate than this benchmark,

To statistically examine the difference between BEAST-GB and this strong benchmark, we used a paired t-test for the prediction errors across all 828 instances. Finally, we generated for each of the benchmark behavioural models in HAB22, a prediction for each instance by averaging all the model's training predictions of that choice task in the train data (that is, in the 15 experimental contexts available for training). A training prediction here is the model's 'prediction' for a participant's choice in a task that was part of the training of the model when it was originally fitted to the data. Hence, these predictions use the entire training data to provide a prediction for out-of-sample behaviour in the test experimental context.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw data for CPC18, as well as processed data for analyses of the previously published datasets (Choices13k and HAB22), are publicly available at https://doi.org/10.17605/OSF.IO/VW2SU.

Code availability

Code for all models and analyses reported in this study is publicly available at https://doi.org/10.17605/OSF.IO/VW2SU.

References

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

2. Tversky, A. & Kahneman, D. Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992).
3. Kahneman, D. & Tversky, A. Prospect theory: an analysis of decision under risk. *Econometrica* **47**, 263–292 (1979).
4. von Neumann, J. & Morgenstern, O. *Theory of Games and Economic Behavior* (Princeton Univ. Press, 1947).
5. Erev, I., Ert, E., Plonsky, O., Cohen, D. & Cohen, O. From anomalies to forecasts: toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychol. Rev.* **124**, 369–409 (2017).
6. He, L., Analytis, P. P. & Bhatia, S. The wisdom of model crowds. *Manag. Sci.* **68**, 3635–3659 (2022).
7. Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D. & Griffiths, T. L. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **372**, 1209–1214 (2021).
8. Altman, A., Bercovici-Boden, A. & Tennenholtz, M. Learning in one-shot strategic form games. In *European Conference on Machine Learning* (eds Fürnkranz, J. et al.) 6–17 (Springer, 2006).
9. Hartford, J. S., Wright, J. R. & Leyton-Brown, K. Deep learning for predicting human strategic behavior. In *Advances in Neural Information Processing Systems* (eds Lee, D. et al.) 2424–2432 (2016).
10. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**, 8–12 (2009).
11. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484 (2016).
12. Peysakhovich, A. & Naecker, J. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *J. Econ. Behav. Organ.* **133**, 373–384 (2017).
13. Fudenberg, D. & Liang, A. Predicting and understanding initial play. *Am. Econ. Rev.* **109**, 4112–4141 (2019).
14. Fudenberg, D., Kleinberg, J., Liang, A. & Mullainathan, S. Measuring the completeness of economic models. *J. Polit. Econ.* **130**, 956–990 (2022).
15. Agrawal, M., Peterson, J. C. & Griffiths, T. L. Scaling up psychology via scientific regret minimization. *Proc. Natl Acad. Sci. USA* **117**, 8825–8835 (2020).
16. Plonsky, O., Erev, I., Hazan, T. & Tennenholtz, M. Psychological forest: predicting human behavior. In *The Thirty-FirstAAAI Conference on Artificial Intelligence* Vol. 31, 656–662 (AAAI Press, 2017).
17. Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J. & Griffiths, T. L. Cognitive model priors for predicting human decisions. In *International Conference on Machine Learning* (eds Chaudhuri, K. & Salakhutdinov, R.) 5133–5141 (PMLR, 2019).
18. Plonsky, O., Apel, R., Erev, I., Ert, E. & Tennenholtz, M. When and how can social scientists add value to data scientists? A choice prediction competition for human decision making. *Open Science Framework* https://doi.org/10.17605/OSF.IO/2X3VT (2018).
19. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
20. Zhou, K., Liu, Z., Qiao, Y., Xiang, T. & Loy, C. C. Domain generalization: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 4396–4415 (2022).
21. Savage, L. J. *The Foundations of Statistics* (John Wiley & Sons, 1954).
22. Allais, M. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econom. J. Econom. Soc.* **21**, 503–546 (1953).
23. Ellsberg, D. Risk, ambiguity, and the Savage axioms. *Q. J. Econ.* **75**, 643–669 (1961).
24. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
25. Dawes, R. M., Faust, D. & Meehl, P. E. Clinical versus actuarial judgment. *Science* **243**, 1668–1674 (1989).
26. Einhorn, H. J. Expert measurement and mechanical combination. *Organ. Behav. Hum. Perform.* **7**, 86–106 (1972).
27. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
28. Thomas, T. et al. Modelling dataset bias in machine-learned theories of economic decision-making. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-023-01784-6 (2024).
29. Shoshan, V., Hazan, T. & Plonsky, O. BEAST-Net: Learning novel behavioral insights using a neural network adaptation of a behavioral model. *Open Science Framework* https://osf.io/kaeny/ (2023).
30. Stewart, N., Reimers, S. & Harris, A. J. L. On the origin of utility, weighting, and discounting functions: how they get their shapes and how to change their shapes. *Manag. Sci.* **61**, 687–705 (2015).
31. Spektor, M. S., Bhatia, S. & Gluth, S. The elusiveness of context effects in decision making. *Trends Cogn. Sci.* **25**, 843–854 (2021).
32. Heilprin, E. & Erev, I. The relative importance of the contrast and assimilation effects in decisions under risk. *J. Behav. Decis. Mak.* **37**, e2408 (2024).
33. Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G. & Scott, C. Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.* **22**, 1–55 (2021).
34. Andrews, I., Fudenberg, D., Liang, A. & Wu, C. The transfer performance of economic models. Preprint at https://arxiv.org/abs/2202.04796 (2022).
35. Dwork, C. et al. The reusable holdout: preserving validity in adaptive data analysis. *Science* **349**, 636–638 (2015).
36. Hofman, J. M. et al. Integrating explanation and prediction in computational social science. *Nature* **595**, 181–188 (2021).
37. Agassi, O. D. & Plonsky, O. The importance of non-analytic models in decision making research: an empirical analysis using BEAST. In *Proc. Annual Meeting of the Cognitive Science Society* (eds Goldwater, M. et al.) 45 (2023).
38. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017).
39. Shafir, S., Reich, T., Tsur, E., Erev, I. & Lotem, A. Perceptual accuracy and conflicting effects of certainty on risk-taking behaviour. *Nature* **453**, 917–920 (2008).
40. Weber, E. U., Shafir, S. & Blais, A.-R. Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychol. Rev.* **111**, 430 (2004).
41. Plonsky, O. & Erev, I. Prediction oriented behavioral research and its relationship to classical decision research. *Open Science Framework* https://doi.org/10.31234/osf.io/7uha4 (2021).
42. d'Eon, G., Greenwood, S., Leyton-Brown, K. & Wright, J. R. How to evaluate behavioral models. In *AAAI Conference on Artificial Intelligence* Vol. 38, 9636–9644 (AAAI Press, 2024).
43. Agassi, O. D. & Plonsky, O. Beyond analytic bounds: re-evaluating predictive power in risky decision models. *Judgm. Decis. Mak.* **19**, e35 (2024).
44. Erev, I., Ert, E., Plonsky, O. & Roth, Y. Contradictory deviations from maximization: environment-specific biases, or reflections of basic properties of human learning? *Psychol. Rev.* **130**, 640–676 (2023).
45. Plonsky, O., Teodorescu, K. & Erev, I. Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychol. Rev.* **122**, 621–647 (2015).
46. Erev, I. & Marx, A. Humans as intuitive classifiers. *Front. Psychol.* **13**, 1041737 (2023).
47. Liu, Y. & Just, A. *SHAPforxgboost: SHAP Plots for 'XGBoost'. R Package Version 0.1.3* (CRAN, 2023).

48. Ert, E. & Erev, I. On the descriptive value of loss aversion in decisions under risk: six clarifications. *Judgm. Decis. Mak.* **8**, 214–235 (2013).
49. Thaler, R. H. & Johnson, E. J. Gambling with the house money and trying to break even: the effects of prior outcomes on risky choice. *Manag. Sci.* **36**, 643–660 (1990).
50. Payne, J. W. It is whether you win or lose: the importance of the overall probabilities of winning or losing in risky choice. *J. Risk Uncertain.* **30**, 5–19 (2005).
51. Birnbaum, M. H. New paradoxes of risky decision making. *Psychol. Rev.* **115**, 463–501 (2008).
52. Barron, G. & Erev, I. Small feedback-based decisions and their limited correspondence to description-based decisions. *J. Behav. Decis. Mak.* **16**, 215–233 (2003).
53. Busemeyer, J. R. & Townsend, J. T. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychol. Rev.* **100**, 432–459 (1993).
54. Diederich, A. & Busemeyer, J. R. Conflict and the stochastic-dominance principle of decision making. *Psychol. Sci.* **10**, 353–359 (1999).
55. Canty, A. & Ripley, B. *boot: Bootstrap R (S-Plus) Functions. R Package Version 1.3-28.1* (CRAN, 2022).
56. Brandstätter, E., Gigerenzer, G. & Hertwig, R. The priority heuristic: making choices without trade-offs. *Psychol. Rev.* **113**, 409–432 (2006).
57. Stewart, N., Chater, N. & Brown, G. D. A. Decision by sampling. *Cogn. Psychol.* **53**, 1–26 (2006).
58. Fiedler, S. & Glöckner, A. The dynamics of decision making in risky choice: an eye-tracking analysis. *Front. Psychol.* **3**, 335 (2012).
59. Rieskamp, J. The probabilistic nature of preferential choice. *J. Exp. Psychol. Learn. Mem. Cogn.* **34**, 1446–1465 (2008).
60. Stewart, N., Hermens, F. & Matthews, W. J. Eye movements in risky choice. *J. Behav. Decis. Mak.* **29**, 116–136 (2016).
61. Pachur, T., Schulte-Mecklenbeck, M., Murphy, R. O. & Hertwig, R. Prospect theory reflects selective allocation of attention. *J. Exp. Psychol. Gen.* **147**, 147–169 (2018).
62. Pachur, T., Mata, R. & Hertwig, R. Who dares, who errs? Disentangling cognitive and motivational roots of age differences in decisions under risk. *Psychol. Sci.* **28**, 504–518 (2017).

Acknowledgements

O.P. thanks O. D. Agassi for help in analysis of some of the curated data. I.E. acknowledges support from the Israel Science Foundation (grant no. 1821/12). M.T. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 740435). D.B., J.C.P., D.R. and T.L.G. have received funding from DARPA (cooperative agreement D17AC00004) and the United States National Science Foundation (grant number 1718550). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

O.P., R.A., E.E., M.T. and I.E. organized CPC18 (O.P., E.E. and I.E. designed the experiments and collected the experimental data; I.E. developed the first baseline model; O.P., R.A. and I.E. programmed the baseline models; O.P. and E.E. managed submissions). D.B., J.C.P., D.R., T.L.G. and S.J.R. submitted the winning model for the first track of CPC18. E.C.C. and J.F.C. submitted the winning model for the second track of CPC18. O.P. performed all post-competition analyses, including analyses of Choices13k and HAB22. O.P. wrote the manuscript, and all authors commented on it.

Competing interests

One of the authors (D.B.) is affiliated with Adobe Research, but his work on this project was done almost exclusively before he had this affiliation. Adobe Research had no role in this project. We declare no other competing interests.

Additional information

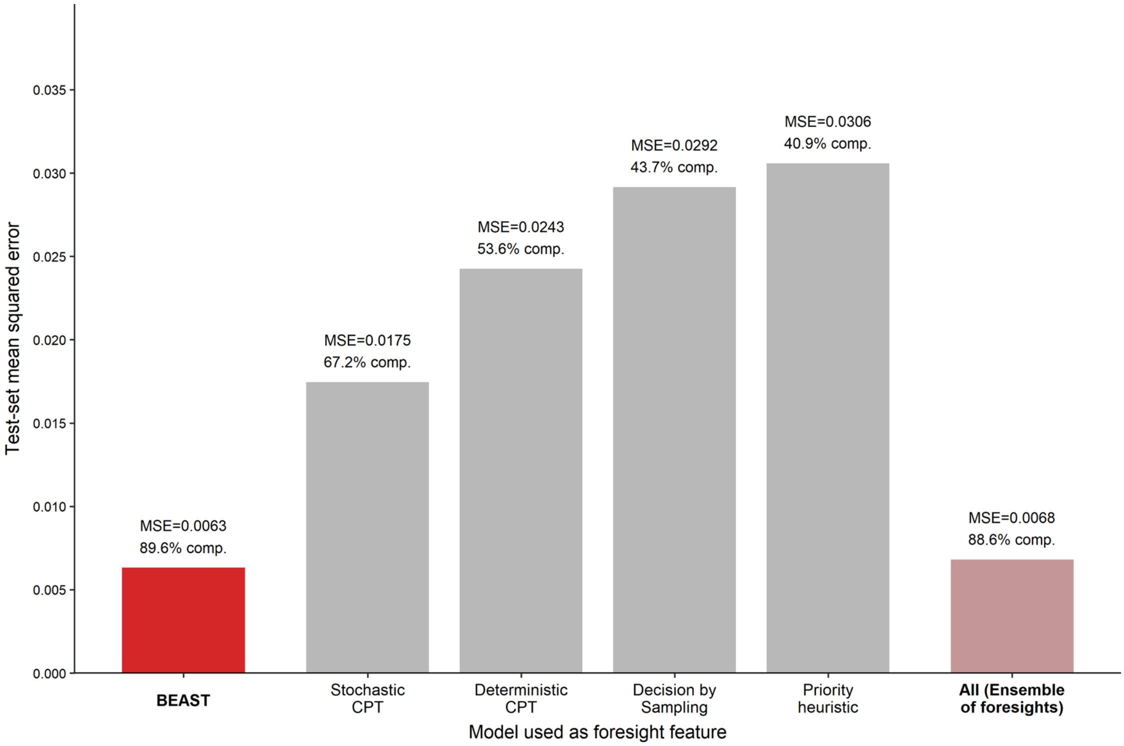
Extended data is available for this paper at https://doi.org/10.1038/s41562-025-02267-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02267-6.

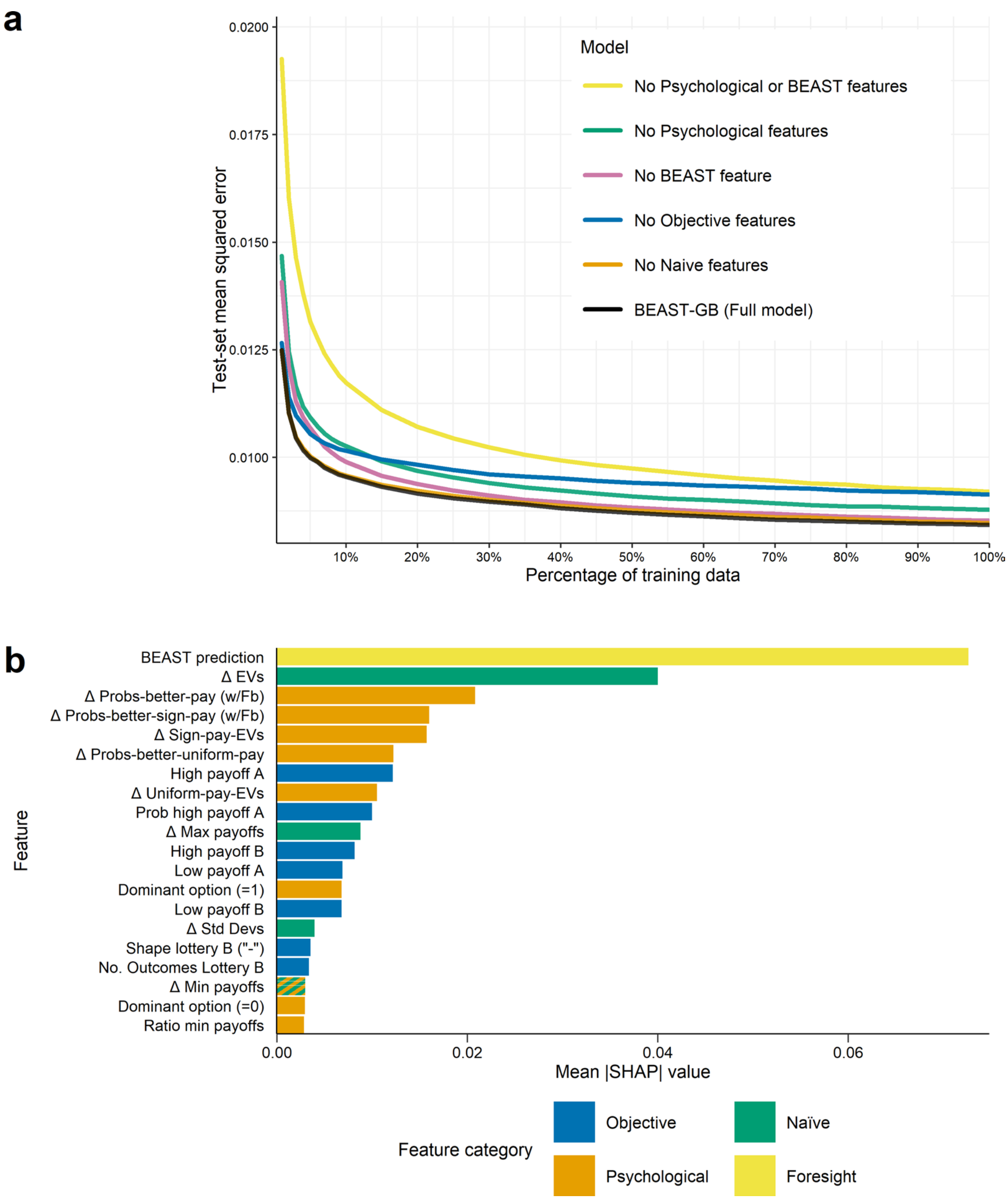
Correspondence and requests for materials should be addressed to Ori Plonsky.

Peer review information *Nature Human Behaviour* thanks Pantelis Analytis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

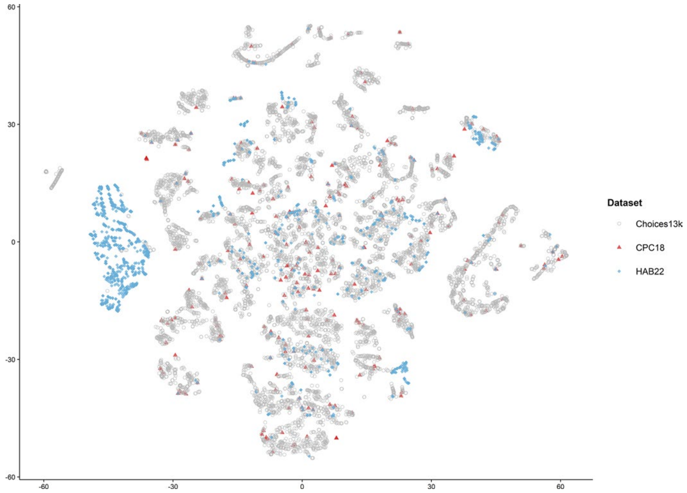
Reprints and permissions information is available at www.nature.com/reprints.



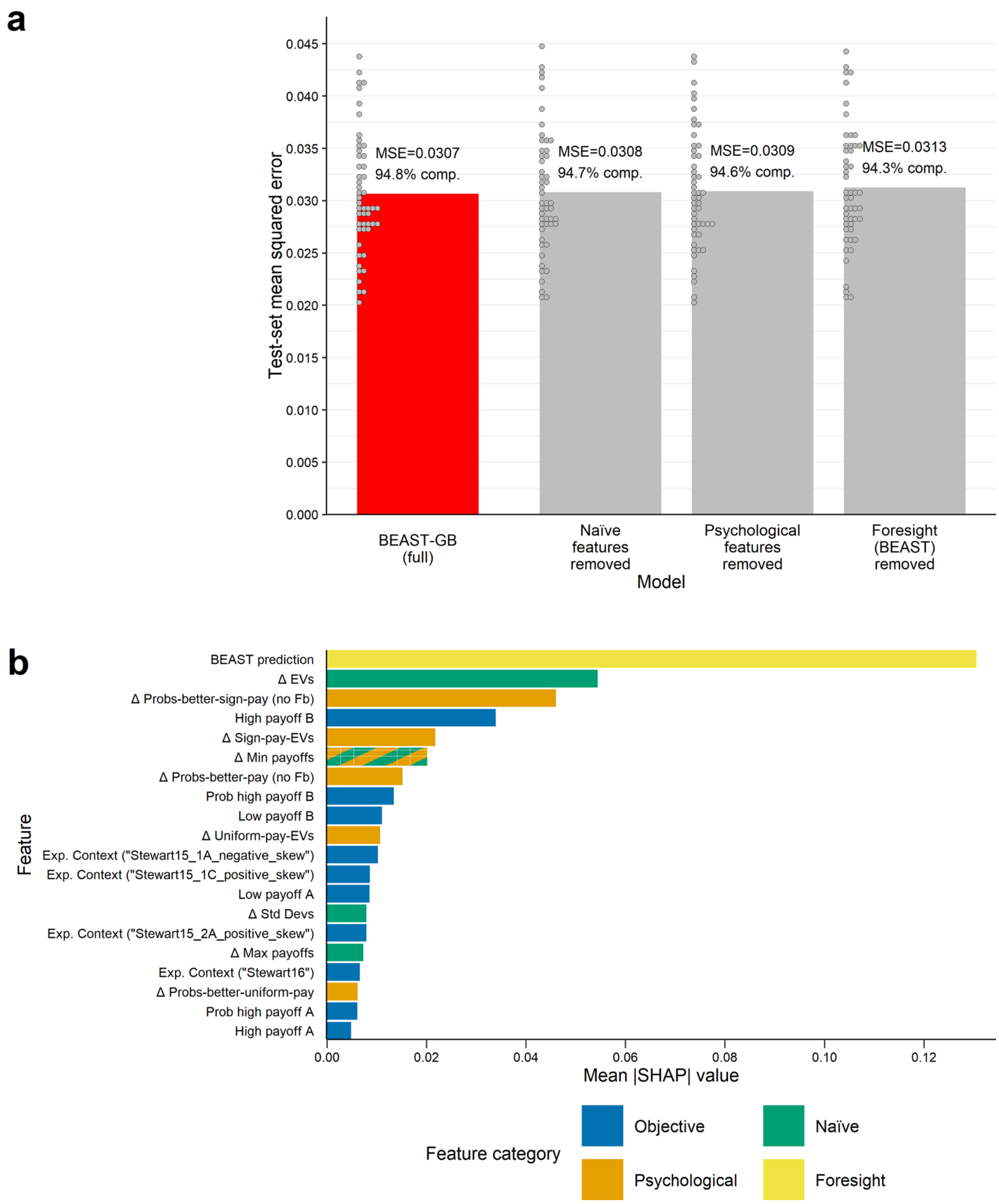
Extended Data Fig. 1 | Comparison of the usefulness of behavioral models as foresight features in CPC18. In each case, we tuned and trained an XGB algorithm using only the objective features (see Table 1) and the prediction of each foresight on CPC18's training data and predicted its test data. Data used was restricted to the subset of CPC18's data that reflects pure decisions under risk (no feedback or ambiguity), implying training on 182 tasks, and testing on 48 tasks. All behavioral models except BEAST were first fitted to the training data independently to provide predictions. BEAST's (red) predictions used the original parameters from CPC15 (Erev et al., 2017). Ensemble of foresights (rosy-brown) uses all five foresights combined. Bars show the single held-out test-set Mean Squared Error (MSE) per model. Completeness (see Methods) computed relative to a naïve baseline (MSE = 0.05095) and irreducible noise limit (MSE = 0.00113).



Extended Data Fig. 2 | Feature importance analyses for Choices13k data. (**a**) Test set performance on Choices13k data when removing different sets of features from BEAST-GB. Data was split to 90% training (8848 tasks) and 10% held-out test data (983 tasks), and models were trained on fixed and increasing proportions of the training data. This process was repeated 50 times, and results reflect the average test set MSE over the n = 50 train-test splits. (**b**) Average absolute SHAP values of BEAST-GB's features in predicting Choices13k test data, by feature category. "Δ Min payoffs" is both a Naïve and a Psychological feature. For clarity, only top 20 features are shown. Feature names and definitions appear in Table 1.



Extended Data Fig. 3 | 2D visualization of all 11,666 choice tasks used in this paper. Each point is a single choice task represented in two dimensions obtained by implementing a t-SNE (t-distributed Stochastic Neighbor Embedding) algorithm on the set of psychological features of each task (see Table 1). Tasks depicted closer together are conceptually more similar than tasks further apart (though the values of the dimensions do not have direct interpretations). Choices13k data appears to cover well the space from which CPC18 data comes from, whereas HAB22 data is different than both.



Extended Data Fig. 4 | Feature importance analyses for HAB22 data. (**a**) HAB22 test set predictive performance of BEAST-GB (red) and variations of it that remove different feature sets. Bars show mean test-set MSE across the 50 nested cross validation folds (n = 50 fold-MSE values). Grey dots form a horizontal dot-histogram of the n = 50 fold-level MSEs (bin = 0.0005) for each model. Completeness (see Methods) computed relative to a naïve baseline (average MSE = 0.1314) and irreducible noise limit (average MSE = 0.0248), in each fold separately, then averaged. (**b**) Average absolute SHAP values of BEAST-GB's features in predicting HAB22's test set, by feature category. "Δ Min payoffs" is both a Naïve and a Psychological feature. For clarity, only top 20 features are shown. Feature names and definitions in Table 1.

nature portfolio

Corresponding author(s): Ori Plonsky

Last updated by author(s): May 28, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For collection of CPC18's original data, the authors used a self-developed software that allows choice between monetary prospects, as presented in Figure 1 of the manuscript. Code of this software was written in Vb.net and is available upon request from the authors.
Data analysis	Analyses was done in R (version 4.4.0) using standard open source packages as listed in the manuscript. Analysis code, including code developed for the models BEAST and BEAST-GB is fully available at https://doi.org/10.17605/OSF.IO/VW2SU

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw data for CPC18, as well as processed data for analyses of the previously published datasets (Choices13k and HAB22) is publicly available at https://doi.org/10.17605/OSF.IO/VW2SU

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	In CPC18, participants self reported their sex but not gender. See below for overall numbers of males and females. We did not analyze any results by sex but the open raw data includes this information for potential future analyses. We were not involved in the data collection in Choices13k or HAB22 data.
Reporting on race, ethnicity, or other socially relevant groupings	We did not collect any such information.
Population characteristics	Participants in CPC18 were (mostly undergraduate) students from either the Technion - Israel Institute of Technology or from the Hebrew University of Jerusalem (half collected in each university). We were not involved in the recruitment of participants in Choices13k or HAB22 data.
Recruitment	In CPC18, Participants were recruited in two ways. Most were recruited from a pool of participants registered for experiments. They received notice on the availability to register for specific time slots and registered to open slots on a first-come-first-served basis. Additional participants were recruited using ads spread around campus. This procedure was known to participants of the competition prior to submission of models. Therefore, any self-selection bias should have no effect on the conclusions of the competition. We were not involved in the recruitment of participants in Choices13k or HAB22 data.
Ethics oversight	Original experiments (CPC18) were approved by the Social and Behavioral Sciences Institutional Review Board in the Technion and by the Ethics Committee for Human Studies at the Faculty of Agriculture, Food, and Environment at the Hebrew University of Jerusalem.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	*Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data exclusions	*Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*
Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This project includes both original collection of data (CPC18 section) and analyses of publicly available datasets (other sections). CPC18 is a quantitative lab study. Data in this paper is split to "train" and "test" sets. Train set is a compound of previously collected data (published elsewhere, see doi:10.1037/rev0000062) and newly collected data (Experiment 1), which in itself is compounded of 2 cohorts ("sets"). Test set includes only newly collected data (Experiment 2), which includes 2 other cohorts ("sets"). The other datasets were collected/assembled by Bourgin et al., 2019 (Choices13k), and by He et al., 2022 (HAB22) and their descriptions are provided in the corresponding papers.
Research sample	Participants in CPC18 were (mostly undergraduate) students from either the Technion - Israel Institute of Technology or from the

Research sample

Hebrew University of Jerusalem (half collected in each university). In Experiment 1, 139/240 were females and the mean age was 24.46 (Range = [18, 37]). In Experiment 2, 141/240 were females and the mean age was 24.7 (Range = [18, 50]). This was a convenience sample which is not representative of the population. The sampling strategy was chosen following the same strategy used in collection of data used in the prior experiments that were also used in this study as training data.

Sampling strategy

In CPC18, Participants were recruited in two ways. Most were recruited from a pool of participants registered for experiments. They received notice on the availability to register for specific time slots and registered to open slots on a first-come-first-served basis. Additional participants were recruited using ads spread around campus.
Sample size was pre-determined based on the sample sizes of previous experiments that were also used in this study as training data. Sample sizes were known to participants of the competition prior to submission. Therefore, there should be no effect for the exact sample size used on the conclusions of the competition.

Data collection

In CPC18, in each experiment, data was collected in sessions of several participants (between 2-6) each. All participants in a session were seated in front of a computer screen by the experimenter. Participants did not communicate to each other and no one else was present during the run. All participants in the same session faced the same set of problems (cohort). A set of problems was assigned to a session before participants arrived in the lab.

Timing

For CPC18, Experiment 1 was run on May-June 2017. Experiment 2 was run on June-July 2018. The gap in collection is in order for the organizers to stay ignorant with respect to the test data for as long as possible.

Data exclusions

No data was excluded from analysis

Non-participation

In CPC18, no participants dropped out or declined to participate.

Randomization

In CPC18, there were no experimental groups. Participants were run in cohorts. Each cohort faced one set of problems run in random order. Each experimental session was assigned a set in advance, so as to balance the number of participants in each set.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.

Research sample

Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

Sampling strategy

Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

Data collection

Describe the data collection procedure, including who recorded the data and how.

Timing and spatial scale

Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Reproducibility

Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

Randomization

Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

Blinding

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work? □ Yes □ No

Field work, collection and transport

Field conditions

Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).

Location

State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).

Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern
☒	□ Plants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

□ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals

For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Reporting on sex

Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
□	□	Public health
□	□	National security
□	□	Crops and/or livestock
□	□	Ecosystems
□	□	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
□	□	Demonstrate how to render a vaccine ineffective
□	□	Confer resistance to therapeutically useful antibiotics or antiviral agents
□	□	Enhance the virulence of a pathogen or render a nonpathogen virulent
□	□	Increase transmissibility of a pathogen
□	□	Alter the host range of a pathogen
□	□	Enable evasion of diagnostic/detection modalities
□	□	Enable the weaponization of a biological agent or toxin
□	□	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

ChIP-seq

Data deposition

□ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

□ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

□ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

□ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

□ All plots are contour plots with outliers or pseudocolor plots.

□ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI □ Used □ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring	*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings	*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*
Effect(s) tested	*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: □ Whole brain □ ROI-based □ Both

Statistic type for inference (See Eklund et al. 2016)	*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*
Correction	*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
□	□ Functional and/or effective connectivity
□	□ Graph analysis
□	□ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*
Graph analysis	*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*
Multivariate modeling and predictive analysis	*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*