Outcome-Irrelevant and State-Independent Learning Mechanisms in Human Reinforcement Learning

Ido Ben-Artzi
Sagol School of Neuroscience; School of Psychological Sciences;
Minducate Science of Learning Research and Innovation Center
Tel-Aviv University
idobenartzi@mail.tau.ac.il

Nitzan Shahar
Sagol School of Neuroscience; School of Psychological Sciences;
Minducate Science of Learning Research and Innovation Center
Tel-Aviv University

Abstract

Humans are adept at associating actions with rewards while considering their state in the environment. However, recent evidence suggests that humans also tend to learn in a *state-independent* manner, resulting in *outcome-irrelevant learning*. This study explores this phenomenon, where individuals form associations between action features and outcomes, even when these associations are known with high certainty to be due to random noise. Using a multi-armed bandit task, we demonstrate that humans tend to rely on a reward-dependent preference for spatial action features in making their decisions, despite knowing these are not predictive of outcomes. Through computational modeling and simulations, we show that this behavior, though sub-optimal in stable environments, may offer adaptive advantages in situations involving unexpected changes. The findings have implications for both understanding human cognition and improving machine learning algorithms by incorporating flexibility in state representations. Our results suggest that humans' predisposition for state-independent learning may reflect an evolved strategy to anticipate environmental variability.

1 Introduction

Humans are highly skilled at inferring the causal structure of their environment, likely more so than any other known living organism [1, 2]. Research has consistently demonstrated humans' ability to identify and act according to latent causal environmental structures which can be learned through a variety of cognitive methods such as trial-and-error value updating [3, 4], model-based reasoning [5, 6, 7, 8], or social interactions [9, 10]. Studies also demonstrated that humans hold in mind sophisticated cognitive causal maps, supported by brain regions such as the orbitofrontal cortex, hippocampus, entorhinal cortex, and striatum [11, 12, 13]. These mental representations enable accurate predictions of the outcome of a specific action based on an inferred state of the environment. Accordingly, most theories propose that human causal learning and decision-making occur in a *state-dependent* manner.

In common RL models, the state representation defines what features of the task are assumed to be predictive of rewards or guide action-outcome associations [14, 15, 16]. Actions may be composed of multiple features, some of which are irrelevant to outcome prediction. Therefore, the cognitive representation of the state must include information about which action features are

NeurIPS 2024 Workshop on Behavioral Machine Learning.

appropriate for credit assignment [17, 4]. If the state representation is inaccurate or cannot be precisely implemented due to resource limitations, an irrelevant feature may mistakenly receive credit [18]. This misattribution can result in credit being generalized to a different action that shares the same irrelevant feature.

Despite a truly sophisticated human capacity for causal inferences, recent striking evidence emerged suggesting that humans also tend to form causal inferences in a state-independent manner that violates the environment's true structure and neglects any accurate causal knowledge the individual might have [19, 20, 21, 22, 23, 24]. Here, we suggest that the human cognitive system continuously tracks and stores temporally adjacent action-outcome associations in a state-independent manner, giving rise to outcome-irrelevant learning. We first review the evidence for this claim and then use a computational model to explore the potential role of this type of learning to guide the development of machine-learning algorithms. We propose that, although not optimal, an outcome-irrelevant policy may be advantageous in environments with frequent unexpected changes. [25, 26, 27, 28].

Empirical evidence for outcome-irrelevant learning in humans. *Outcome-irrelevant learning refers to assigning value to action features that are reliably known to lack outcome relevance.* A clear example of this is seen in [19], where participants performed a multi-armed bandit task with four cards. Specifically, the same pairs of cards were offered (A vs B or C vs D, see Figure 1). Each card had a set probability of reward that drifted slowly across trials requiring participants to continuously keep track of action-outcome associations to make optimal card selections. Importantly, the cards were randomly assigned to left/right response keys each trial, making the spatial locations/response keys irrelevant to predicting outcomes — a fact emphasized to participants through instructions and a verbal quiz. Thus, participants could learn from both instructions and their own experience that only the cards' identity is relevant for reward prediction. In a critical analysis [19] predicted the selection of C/D cards based on the reward obtained in the previous trial where A/B cards were offered (see Figure 1). Despite participants being fully aware that there is no dependency between the trials they were more likely to use the same response key in the C/D trial after a reward in the A/B trial compared with an unrewarded trial.

A few studies replicated and extended the findings regarding outcome-irrelevant learning in humans. First, outcome-irrelevant learning persists over time and practice, both within and between sessions [22]. Importantly, outcome-irrelevant learning does not stem from a lack of comprehension or belief in the correctness of the instructions. Findings demonstrate that even when the state representation was presented clearly and concretely outcome-irrelevant learning remained [29, 24]. Moreover, in a recent work, outcome-irrelevant learning was found across modalities, showing that humans assign value to both visual and motor/spatial representation when these are known to be outcome-irrelevant [30]. The effect persisted even among participants who rated their belief in the instructions as maximal and explicitly indicated no influence of the outcome-irrelevant feature on their choices [29]. Finally, a stronger human tendency for outcome-irrelevant learning is associated with lower cognitive control estimates including model-based planning [23], working memory capacity, and IQ [19, 20].

2 Results

We begin by outlining the behavioral task, computational modeling, and empirical parameter estimation based on human behavior. Next, we use these empirical parameters to drive a reinforcement learning simulation, illustrating how the human tendency for outcome-irrelevant learning can augment artificial model behavior under specific conditions.

Participants and data collection. We reanalyze the behavior of 178 Prolific workers (age mean = 26.1, range 18 to 51; 101 males, 76 females, 1 other) completing the task online in return for monetary compensation (see OSF repository).

Behavioral Task. Participants completed a modified version of a four-armed bandit task (see Figure 1). In each trial, two out of four cards (i.e., bandit arms) are offered for choice, with the allocation determined randomly. Cards probabilistically lead to binary outcomes (later translated to money bonuses) according to a random walk specific for each arm. Importantly, human participants are explicitly made aware of the fact that only the card itself influences reward probability and not its location. Thus, they know with high certainty that the card's location (i.e., whether it appears on the left or right of the screen) is random and therefore should not be learned. This fact is crucial as it renders any credit assignment to the location/response key as outcome-irrelevant.

Model-agnostic results. We find a replicable human tendency to assign credit to outcome-irrelevant locations. This is indicated by analyzing behavior on trials where two different cards are offered than in the previous trial (see trial n+1 in Figure 1A). A hierarchical logistic Bayesian regression shows that individuals are more likely to stay with their response key selection if the previous offer was rewarded (51%) vs. unrewarded (45%; posterior median=.23, $\text{HDI}_{95\%}$ between .12 and .34; probability of direction ∼100%; see Figure 1B-C)). We further found that this effect is the same irrespective of whether the reward condition is a win (+1/0) or loss (0/-1) block and whether staying was the accurate choice (see Figure 1B-C; Figure S1).

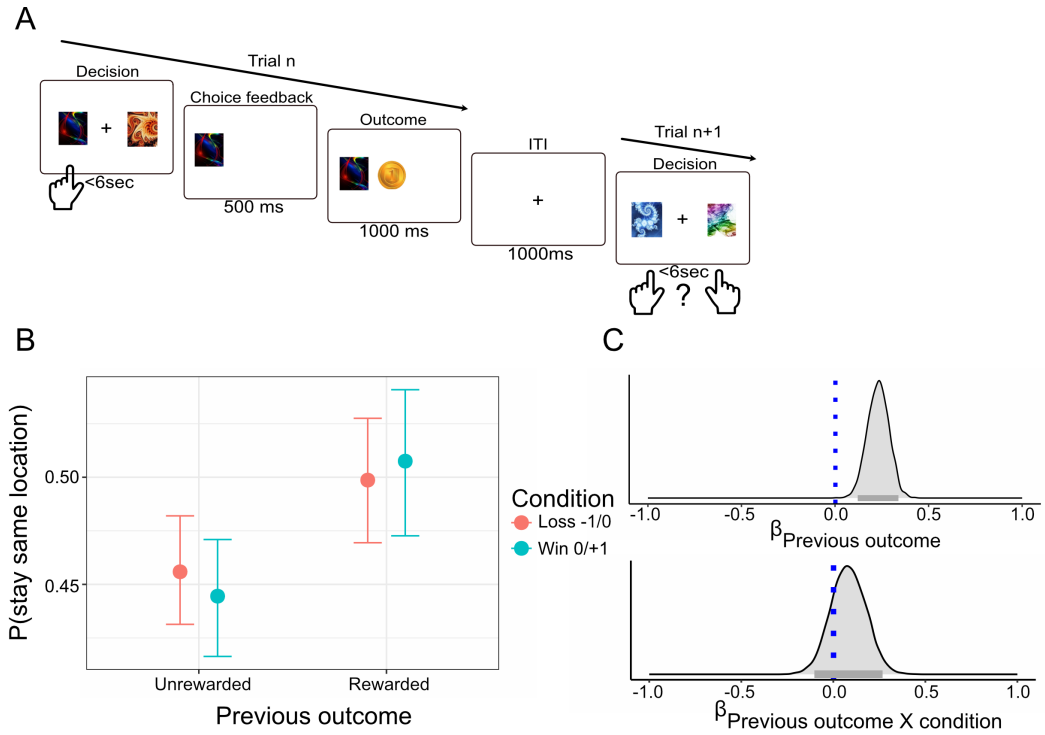

Figure 1: **Trial sequence and main behavioral result.** (A) Illustration of a trial sequence. Participants completed a four-armed bandit task. In each trial, two cards (of four) were randomly offered by the computer for participants' selection. Participants were explicitly aware that only the card itself and not the card's location (which was randomly determined by the computer) influences reward probability. (B) The empirical probability of choosing the same location as a function of outcome in the previous trial (where two different cards were offered for choice) and win/loss condition. Participants were more likely to choose the same location after a rewarded (51%) compared to an unrewarded (45%) trial. This was true for both win blocks (where rewards were +1 or 0) and loss blocks (where rewards were 0 or -1). (C) The posterior distributions for the influence of the previous outcome (top) and the interaction with condition (bottom) on choosing the same location (the blue line is the null point, and the gray line indicates $\text{HDI}_{95\%}$).

Computational model. To describe the trial-by-trial value updating process we use a simple feature-based reinforcement learning model [31, 32, 33]. Action feature values are updated using a temporal difference learning algorithm (There are separate Q-values for the relevant visual feature and irrelevant spatial-motor feature). On each trial, the reward is used to calculate the prediction errors of each action feature (Eq.1; denoted δ_f). These prediction errors are then used to update the values of the chosen action features where α is the learning rate (free-parameter; Eq.2).

$$\delta_f = (r_t - Q_f) \tag{1}$$

$$Q_f = Q_f + \alpha \cdot \delta_f \tag{2}$$

The agent can follow the clearly defined state representation (*instructed_weights*; where a weight of 1 is given for the relevant feature and 0 for any irrelevant feature; Eq. 3) or ignore it to treat each feature in the same manner (*uniform_weights*; 1 divided by the number of features for each action

feature; Eq. 5).

$$\textit{instructed_weights} = \begin{cases} 1 & \text{if feature is relevant} \\ 0 & \text{if feature is irrelevant} \end{cases} \tag{3}$$

$$\mathbf{W} \in \mathbb{R}^{N_{\text{features}}} \tag{4}$$

$$\textit{uniform_weights} = \left(\frac{1}{\text{Nfeatures}} \right)_{\mathbf{1} \times \text{Nfeatures}} \tag{5}$$

Higher λ values lead to a greater tendency to rely on the instructed state representation. On the contrary, lower λ values indicate a diminished focus on the instructed state representation. Note that λ is scaled using a logistic transformation to be within the range of 0 to 1 (free parameter; Eq. 6).

$$W = \lambda \cdot \textit{instructed_weights} + (1 - \lambda) \cdot \textit{uniform_weights} \tag{6}$$

The decision policy of this model includes value integration according to feature-specific decision weights (w_f). For each possible action (e.g., choosing the specific card offered on the left), the values of applicable action features (e.g., the value of the left location and the value of the specific card on the left) are combined (Eq. 7).

$$Q_{\text{net}} = \sum_f (w_f \cdot Q_f) \tag{7}$$

The policy is determined by a softmax function transforming net values into action probabilities and includes an inverse noise parameter β controlling for random exploration tendencies (free parameter; Eq. 8).

$$p(\text{choice}_i) = \frac{e^{\beta \cdot Q_{\text{net } i}}}{\sum_j e^{\beta \cdot Q_{\text{net } j}}} \tag{8}$$

Empirical parameter estimation. We used a hierarchical Bayesian reinforcement learning model in Stan to derive individual fits for each parameter based on empirical data (see Subsection 4.1 and Figure S2 for parameter recovery results;[34, 35]). The empirical population-level posterior estimates for each of the three parameters in our model were as follows: $\alpha \sim \mathcal{N}(.4, .18)$; $\beta \sim \mathcal{N}(2.3, .82)$; $\lambda \sim \mathcal{N}(1.2, 2)$ (see Figure S3 for transformed λ fits). The adequacy of our model fit is further demonstrated by the ability of the λ parameter to account for individual differences in the observed behavioral effect (see Figure S4).

Simulation. To illustrate the potential benefit of outcome-irrelevant learning we simulated 500 agents who completed each 100 trials of the task. Importantly, after 50 trials we introduced an unexpected and uncued shift in the true environment's representation (see Figure 2A for illustration). Parameters for the agents were sampled from the posterior empirical fit distribution. We used hierarchical Bayesian regression to predict the trial's accuracy (whether the higher or lower expected value arm was chosen) based on the agents' λ parameter and task phase (i.e., before or after the shift). We show that while outcome-irrelevant learning is detrimental before the shift as evidenced by the association of lower λ with lower accuracy (posterior median=.17, $\text{HDI}_{95\%}$ between .04 and .31; probability of direction $\sim$100%; see Figure 2B), it also allows agents to better adapt to change as evident by a higher accuracy following the shift (interaction posterior median=-.36, $\text{HDI}_{95\%}$ between -.55 and -.17; probability of direction $\sim$100%; see Figure 2B). Importantly, agents who blindly followed the state representation (i.e., λ=1) were doomed to random responses following the shift, thus indicating the potential adaptivity of retaining a degree of flexibility in state representation.

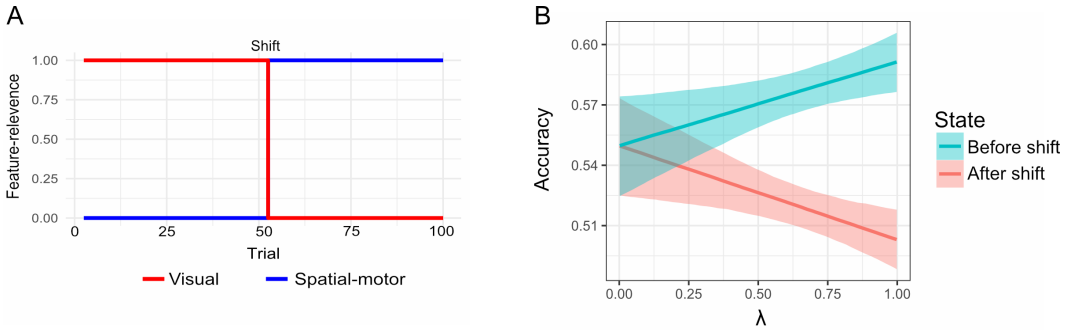

Figure 2: **The accuracy of outcome-irrelevant learning agents is higher after an uncued shift.** (A) We simulated 500 agents performing a multi-armed bandit task using empirical parameters. After 50 out of 100 trials, we introduced an unexpected and uncued shift in feature relevance. (B) We found that outcome-irrelevant learners were better able to adapt to the shift, as indicated by higher accuracy levels following it. Furthermore, the variance in accuracy (and thereby reward) was smaller for lower λ agents.

3 Discussion

Current findings suggest that humans tend to overlook knowledge about state representation. This finding is supported by results from various tasks and research teams [36, 20, 21, 22, 19, 24]. The established preferences are implicit and persist regardless of belief in the instructions or self-reported intentions. Importantly, unlike previously described sequential effects, this behavior is not driven by random or perseverative tendencies but is instead dependent on rewards [28, 37]. Future studies may explore how instructed influence interacts with experiential feature-weight learning [38, 39]. Perhaps, the influence of an instructed state representation on the RL cognitive system is limited. This hypothesis is further supported by the negative correlations between working memory abilities or model-based learning in the two-step task and state-independent learning [19, 22].

In the multi-armed bandit task, we assume participants represent a static single state; thus, any outcome-irrelevant learning is incorporated within this state representation. However, we also observed outcome-irrelevant learning when participants represented multiple states. In one case, participants completed a task where the state representation switched between trials; in another, they completed a two-step task where rewards from the second stage influenced location choices in the first stage [30, 22]. Thus, outcome-irrelevant learning could occur due to an incorrect representation of a single-state environment but also affects choices independently of the state in multi-state environments.

At first glance, humans' tendency for outcome-irrelevant and state-independent learning appears to contradict the typical effectiveness associated with human cognition. However, recent research suggests evaluating human behavior by reasonableness instead of optimality [18, 40, 26, 41]. Relatedly, we demonstrate that outcome-irrelevant learning, while detrimental in a stable environment, facilitates better adaptation to unexpected changes. We, therefore, speculate that this form of redundant learning arises from humans' underlying assumption that the environment may change unexpectedly [27]. Bet-hedging is a prevalent biological phenomenon in which organisms accept a short-term loss to ensure long-term prosperity [42, 43, 44, 45]. Formally, this can be understood as employing a strategy that sacrifices higher immediate expected fitness to reduce the variance of future outcomes. This aligns with our results, where agents with lower λ earn less before the change but experience a smaller variance in outcomes overall. Other normative interpretations for this human tendency are possible. For example, the effort required to suppress outcome-irrelevant learning may overcome the extra potential reward it promises [46, 47]. Recently, [48] stressed the importance of accounting for human decision-making biases in improving algorithm design. Recognizing the substantial individual differences in outcome-irrelevant learning could provide machine learning models with a more nuanced perspective, potentially reducing model misspecification [49, 50]. Future ML research could further study how models could benefit from outcome-irrelevant and state-independent learning.

4 SI

4.1 Parameter Recovery Results

The model has three population-level free parameters (α, β, λ) along with their associated random effect parameters. Parameters from data simulated using this model were highly recoverable (see Figure S2; [35]). This establishes the appropriateness of using the model when fitting empirical data, as it confirms our ability to estimate parameters from choice data with high confidence.

4.2 Model fit results

Fitting empirical choice data using the model allowed us to estimate each human participant's (λ) parameter which controls the decision weights regulation (see Eq. 7). Using the fitted values we could show that the behavioral signature depends on this model-derived parameter for both simulated and empirical data (see Figure S4A). Further, we show that higher lambda values are associated with greater accuracy (defined as choosing the higher EV arm; (see Figure S4B)).

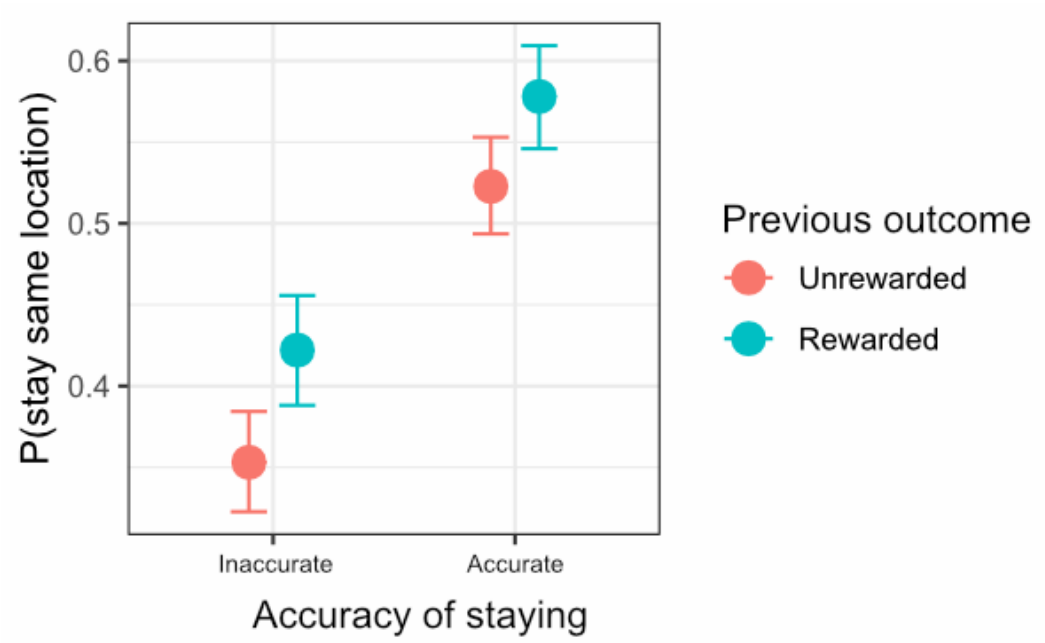

Figure S1: **Participants repeat location after rewarded trials regardless of accuracy.** We further analyzed whether participants showed a different influence by the previous outcome based on the accuracy of choosing the same location. Accuracy was defined as choosing the card with the higher current expected value. Thus, accurately staying is when the better card appeared on the previously selected location and inaccurate staying is when choosing the better card required switching locations. We found the effect of the previous outcome to be the same in these two conditions (posterior median=-.07, $\text{HDI}_{95\%}$ between -.29 and .15; probability of direction=73.5%).

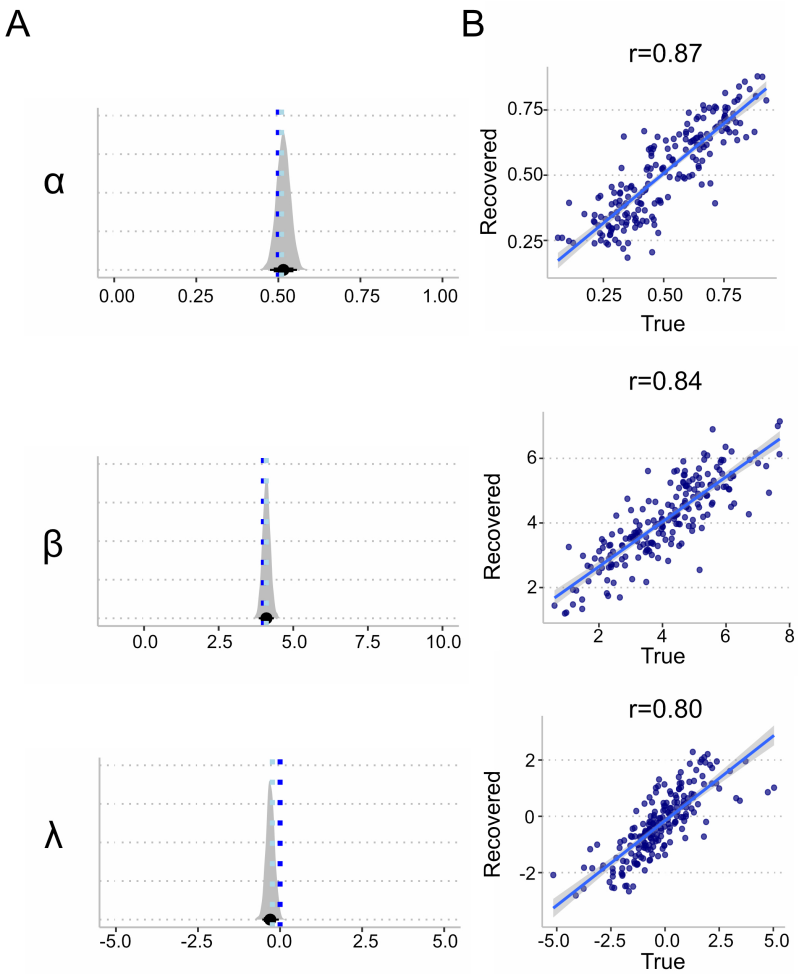

Figure S2: **Parameter recovery of the decision modulation model.** The left column (A) represents population parameter recovery, including the posterior parameter distribution (grey), the value of the true latent population parameter (blue dashed line), and the empirical sample mean (cyan dashed line). The right column (B) refers to individual parameter recovery, showing a strong correlation between simulated individual parameters and recovered ones. Overall, we found great parameter recovery for all parameters.

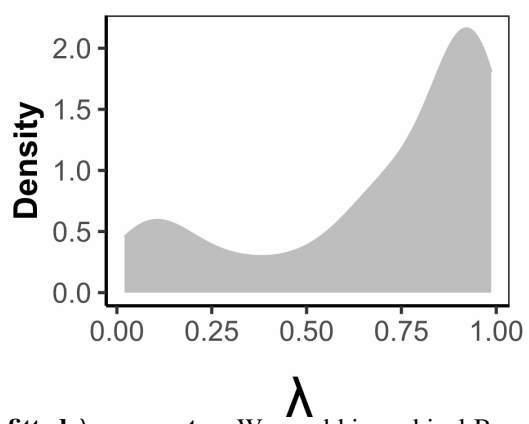

Figure S3: **Empirically fitted λ parameter.** We used hierarchical Bayesian modeling in stan to (Carpenter et al., 2017) estimate for each participant their best-fitting λ parameter. Results show substantial individual differences in the degree to which outcome-irrelevant information affected choices.

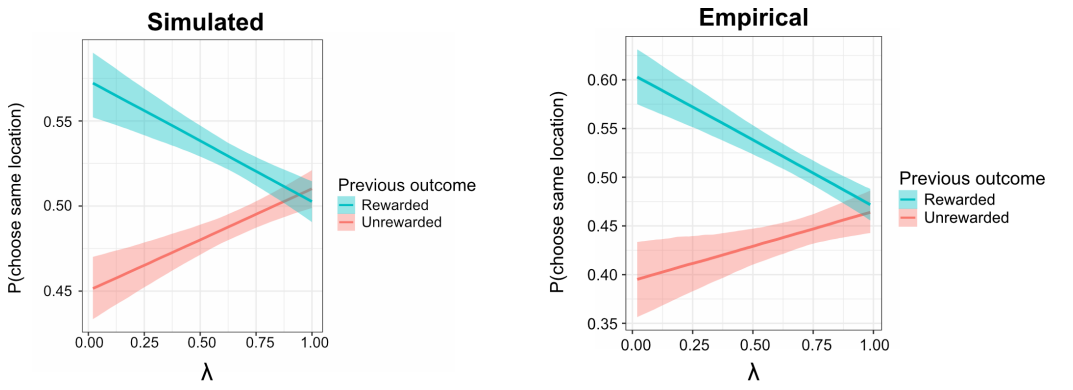

Figure S4: **Simulated and Empirical fit of the model to data.** We show that the main behavioral signature indicating outcome-irrelevant decision-making (see Figure 1) is well-captured by our computational model. The λ parameter was fitted based on empirical data using hierarchical Bayesian modeling in stan [51]. These fitted estimations were then used as an additional predictor in a hierarchical Bayesian logistic regression where the previous outcome in trial n (rewarded versus unrewarded) predicted the tendency to choose in trial n + 1 the same location as in trial n.

References

[1] Aaron P. Blaisdell, Kosuke Sawa, Kenneth J. Leising, and Michael R. Waldmann. Causal Reasoning in Rats. Science, 311(5763):1020–1022, February 2006. Publisher: American Association for the Advancement of Science.

[2] Michael R. Waldmann, York Hagmayer, and Aaron P. Blaisdell. Beyond the Information Given: Causal Models in Learning and Reasoning. Curr Dir Psychol Sci, 15(6):307–311, December 2006. Publisher: SAGE Publications Inc.

[3] Samuel J Gershman, Kenneth A Norman, and Yael Niv. Discovering latent causes in reinforcement learning. Current Opinion in Behavioral Sciences, 5:43–50, October 2015.

[4] Yael Niv. Learning task-state representations. Nat Neurosci, 22(10):1544–1553, October 2019.

[5] Nathaniel D. Daw, Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron, 69(6):1204–1215, March 2011.

[6] Bradley B Doll, Dylan A Simon, and Nathaniel D Daw. The ubiquity of model-based reinforcement learning. Current Opinion in Neurobiology, 22(6):1075–1081, December 2012.

[7] Carolina Feher Da Silva and Todd A. Hare. Humans primarily use model-based inference in the two-stage task. Nat Hum Behav, 4(10):1053–1066, July 2020.

[8] Carolina Feher Da Silva, Gaia Lombardi, Micah Edelson, and Todd A. Hare. Rethinking model-based and model-free influences on mental effort and striatal prediction errors. Nat Hum Behav, 7(6):956–969, April 2023.

[9] Albert Bandura. Social learning through imitation. In Nebraska Symposium on Motivation, 1962, pages 211–274. Univer. Nebraska Press, Oxford, England, 1962.

[10] Rachit Dubey, Hermish Mehta, and Tania Lombrozo. Curiosity Is Contagious: A Social Influence Intervention to Induce Curiosity. Cognitive Science, 45(2):e12937, February 2021.

[11] Mona M Garvert, Raymond J Dolan, and Timothy EJ Behrens. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. eLife, 6:e17086, April 2017. Publisher: eLife Sciences Publications, Ltd.

[12] Rani Moran, Peter Dayan, and Raymond J. Dolan. Human subjects exploit a cognitive map for credit assignment. PNAS, 118(4), January 2021. ISBN: 9782016884119 Publisher: National Academy of Sciences Section: Biological Sciences.

[13] Nicolas W. Schuck, Ming Bo Cai, Robert C. Wilson, and Yael Niv. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. Neuron, 91(6):1402–1412, September 2016.

[14] P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci, 16(5):1936–1947, March 1996.

[15] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. Science, 275(5306):1593–1599, March 1997.

[16] R. E. Suri and W. Schultz. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience, 91(3):871–890, 1999.

[17] Nadav Amir, Yael Niv, and Angela Langdon. States as goal-directed concepts: an epistemic approach to state-representation learning, January 2024. arXiv:2312.02367.

[18] Thomas L. Griffiths, Falk Lieder, and Noah D. Goodman. Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. Topics in Cognitive Science, 7(2):217–229, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12142.

[19] Ido Ben-Artzi, Roy Luria, and Nitzan Shahar. Working memory capacity estimates moderate value learning for outcome-irrelevant features. Sci Rep, 12(1):19677, November 2022.

[20] Azadeh Nazemorroaya, Dan Bang, and Peter Dayan. State-Independent and State-Dependent Learning in a Motivational Go/NoGo task. Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0), 2024.

[21] Milena Rmus, Amy Zou, and Anne GE Collins. Choice Type Impacts Human Reinforcement Learning. Journal of Cognitive Neuroscience, 35(2):314–330, 2023. Publisher: MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA . . . .

[22] Nitzan Shahar, Rani Moran, Tobias U. Hauser, Rogier A. Kievit, Daniel McNamee, Michael Moutoussis, NSPN Consortium, Raymond J. Dolan, Edward Bullmore, Raymond Dolan, Ian Goodyer, Peter Fonagy, Peter Jones, Michael Moutoussis, Tobias Hauser, Sharon Neufeld, Rafael Romero-Garcia, Michelle St Clair, Petra Vértes, Kirstie Whitaker, Becky Inkster, Gita Prabhu, Cinly Ooi, Umar Toseeb, Barry Widmer, Junaid Bhatti, Laura Villis, Ayesha Alrumaithi, Sarah Birt, Aislinn Bowler, Kalia Cleridou, Hina Dadabhoy, Emma Davies, Ashlyn Firkins, Sian Granville, Elizabeth Harding, Alexandra Hopkins, Daniel Isaacs, Janchai King, Danae Kokorikou, Christina Maurice, Cleo McIntosh, Jessica Memarzia, Harriet Mills, Ciara O'Donnell,

Sara Pantaleone, Jenny Scott, Pasco Fearon, John Suckling, Anne-Laura Van Harmelen, and Rogier Kievit. Credit assignment to state-independent task representations and its relationship with model-based decision making. Proc. Natl. Acad. Sci. U.S.A., 116(32):15871–15876, August 2019.

[23] Nitzan Shahar, Tobias U. Hauser, Rani Moran, Michael Moutoussis, NSPN consortium, Principal investigators, Edward Bullmore, Raymond J. Dolan, Ian Goodyer, Peter Fonagy, Peter Jones, NSPN (funded) staff, Michael Moutoussis, Tobias Hauser, Sharon Neufeld, Rafael Romero-Garcia, Michelle St Clair, Petra Vértes, Kirstie Whitaker, Becky Inkster, Gita Prabhu, Cinly Ooi, Umar Toseeb, Barry Widmer, Junaid Bhatti, Laura Villis, Ayesha Alrumaithi, Sarah Birt, Aislinn Bowler, Kalia Cleridou, Hina Dadabhoy, Emma Davies, Ashlyn Firkins, Sian Granville, Elizabeth Harding, Alexandra Hopkins, Daniel Isaacs, Janchai King, Danae Kokorikou, Christina Maurice, Cleo McIntosh, Jessica Memarzia, Harriet Mills, Ciara O'Donnell, Sara Pantaleone, Jenny Scott, Beatrice Kiddle, Ela Polek, Affiliated scientists, Pasco Fearon, John Suckling, Anne-Laura Van Harmelen, Rogier Kievit, Sam Chamberlain, Edward T. Bullmore, and Raymond J. Dolan. Assigning the right credit to the wrong action: compulsivity in the general population is associated with augmented outcome-irrelevant value-based learning. Transl Psychiatry, 11(1):564, November 2021.

[24] Yael Troudart and Nitzan Shahar. Formation of non-veridical action-outcome associations following exposure to threat-related cues. Emotion, 23(7):2094–2099, 2023. Place: US Publisher: American Psychological Association.

[25] Falk Lieder and Thomas L. Griffiths. Strategy selection as rational metareasoning. Psychological Review, 124(6):762–794, 2017. Place: US Publisher: American Psychological Association.

[26] Jens Koed Madsen, Lee de Wit, Peter Ayton, Cameron Brick, Laura de Moliere, and Carla J. Groom. Behavioral science should start by assuming people are reasonable. Trends in Cognitive Sciences, 28(7):583–585, July 2024. Publisher: Elsevier.

[27] Arne Öhman. Has evolution primed humans to "beware the beast"? Proc. Natl. Acad. Sci. U.S.A., 104(42):16396–16397, October 2007.

[28] Angela J. Yu and Jonathan D. Cohen. Sequential effects: Superstition or rational behavior? Adv Neural Inf Process Syst, 21:1873–1880, 2008.

[29] Ido Ben-Artzi and Nitzan Shahar. The influence of story-based instructions on credit assignment to outcome-irrelevant action features, July 2024.

[30] Nitzan Shahar, Philipp Schwartenbeck, and Raymond J Dolan. Examining state-shielding during value updating and decision-making using a cued multi-dimensional bandit task., 2024.

[31] Rachel S. Lee, Yotam Sagiv, Ben Engelhard, Ilana B. Witten, and Nathaniel D. Daw. A feature-specific prediction error model explains dopaminergic heterogeneity. Nat Neurosci, 27(8):1574–1586, August 2024.

[32] Pantelis Pipergias Analytis, Maarten Speekenbrink, and Hrvoje Stojic. Human behavior in contextual multi-armed bandit problems: 37th Annual Meeting of the Cognitive Science Society. Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, 1:2290–2295, 2015. Publisher: Cognitive Science Society.

[33] Hrvoje Stojić, Eric Schulz, Pantelis P. Analytis, and Maarten Speekenbrink. It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. Journal of Experimental Psychology: General, 149(10):1878–1907, October 2020.

[34] Jonah Gabry. cmdstanr: R Interface to'CmdStan', 2021.

[35] Robert C Wilson and Anne GE Collins. Ten simple rules for the computational modeling of behavioral data. eLife, 8:e49547, November 2019. Publisher: eLife Sciences Publications, Ltd.

[36] Nir Moneta, Mona M. Garvert, Hauke R. Heekeren, and Nicolas W. Schuck. Task state representations in vmPFC mediate relevant and irrelevant value signals and their behavioral influence. Nat Commun, 14(1):3156, May 2023. Number: 1 Publisher: Nature Publishing Group.

[37] Kevin J. Miller, Amitai Shenhav, and Elliot A. Ludvig. Habits without values. Psychological Review, 126(2):292–311, 2019. Place: US Publisher: American Psychological Association.

[38] Bradley B. Doll, W. Jake Jacobs, Alan G. Sanfey, and Michael J. Frank. Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. Brain Research, 1299:74–94, November 2009.

[39] Yael Niv, Reka Daniel, Andra Geana, Samuel J. Gershman, Yuan Chang Leong, Angela Radulescu, and Robert C. Wilson. Reinforcement learning in multidimensional environments relies on attention mechanisms. J. Neurosci., 35(21):8145–8157, May 2015. Publisher: Society for Neuroscience Section: Articles.

[40] Falk Lieder, Amitai Shenhav, Sebastian Musslick, and Thomas L. Griffiths. Rational metareasoning and the plasticity of cognitive control. PLOS Computational Biology, 14(4):e1006043, April 2018. Publisher: Public Library of Science.

[41] Herbert A. Simon. A Behavioral Model of Rational Choice. The Quarterly Journal of Economics, 69(1):99–118, 1955. Publisher: Oxford University Press.

[42] Luiza P. Morawska, Jhonatan A. Hernandez-Valdes, and Oscar P. Kuipers. Diversity of bet-hedging strategies in microbial communities-Recent cases and insights. WIREs Mech Dis, 14(2):e1544, March 2022.

[43] Tom Philippi and Jon Seger. Hedging one's evolutionary bets, revisited. Trends in Ecology & Evolution, 4(2):41–44, February 1989.

[44] Si Tang, Yaqing Liu, Jianming Zhu, Xueyu Cheng, Lu Liu, Katrin Hammerschmidt, Jin Zhou, and Zhonghua Cai. Bet hedging in a unicellular microalga. Nat Commun, 15(1):2063, March 2024. Publisher: Nature Publishing Group.

[45] Ann T Tate and Jeremy Van Cleve. Bet-hedging in innate and adaptive immune systems. Evolution, Medicine, and Public Health, 10(1):256–265, January 2022.

[46] R. Frömer and A. Shenhav. Filling the gaps: Cognitive control as a critical lens for understanding mechanisms of value-based decision-making. Neuroscience & Biobehavioral Reviews, 134:104483, March 2022.

[47] Amitai Shenhav, Matthew M. Botvinick, and Jonathan D. Cohen. The expected value of control: An integrative theory of anterior cingulate cortex function. Neuron, 79(2):217–240, July 2013.

[48] Carey K. Morewedge, Sendhil Mullainathan, Haaya F. Naushan, Cass R. Sunstein, Jon Kleinberg, Manish Raghavan, and Jens O. Ludwig. Human bias in algorithm design. Nat Hum Behav, 7(11):1822–1824, November 2023.

[49] Jeffrey M. Beck, Wei Ji Ma, Xaq Pitkow, Peter E. Latham, and Alexandre Pouget. Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. Neuron, 74(1):30–39, April 2012.

[50] Yoav Ger, Moni Shahar, and Nitzan Shahar. Using recurrent neural network to estimate irreducible stochasticity in human choice-behavior. eLife, 13, January 2024. Publisher: eLife Sciences Publications Limited.

[51] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. *Stan* : A Probabilistic Programming Language. J. Stat. Soft., 76(1), 2017.