Separable Contributions of Value and Choice to Policy Learning

Maximilian Mittenbühler (maximilian.mittenbuehler@uni-tuebingen.de)[1, 2], Xingrui Gu[3] & Anne G. E. Collins[2, 4]

[1]Department of Computer Science, University of Tübingen, Tübingen, Germany
[2]Department of Psychology, University of California, Berkeley, Berkeley, CA, USA
[3]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA
[4]Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, CA, USA

Abstract

Reinforcement learning (RL) models have been highly successful in describing human behavior. However, in most experiments, reward history is highly correlated with choice history. Consequently, behavior could be partially driven by a value-free habitual learning (HL) mechanism but be misinterpreted as RL. Do people select actions they have chosen frequently in the past (HL), or actions that have been rewarding in the past (RL)? To disentangle the influence of value-based RL and value-free HL, we designed two experiments that unconfound reward and choice histories. Behavioral and modeling results suggest that both value-based and value-free mechanisms contribute to test phase choice. Their relative contributions differ in the two experiments, suggesting that different environments may recruit the two processes to different degrees. Our results highlight the existence of partially redundant processes that are often difficult to disambiguate, and call for careful consideration of underlying processes when interpreting experimental results.

Keywords: reinforcement learning, habits, value-free learning

Introduction

Human learning exhibits signatures of both outcome-dependence and -independence. Outcome-dependent, or *value-based*, behavior occurs when an action is chosen because it has been rewarded in the past, a mechanism explicitly defined in model-free reinforcement learning (RL; Sutton and Barto, 2018); for example, selecting a restaurant because of its past high quality. Outcome-independent, or *value-free* behavior, on the other hand, is manifested if an action is chosen simply because it has been chosen in the past, which is typically associated with habitual learning (HL; Watson et al., 2022; Wood and Rünger, 2016); for example, selecting a restaurant because it is the usual one. These two modes of decision making are fundamentally distinct, resulting in different behaviors and implications (Dezfouli and Balleine, 2012; Miller et al., 2019; Wood et al., 2022). Understanding which mode drives behavior is therefore crucial for making further predictions about the generality of learned associations or the way humans react to erroneous or devalued actions. Under most typical experimental designs, however, the choice history of actions, supporting HL, closely aligns with the reward history, supporting RL, because participants choose more rewarding outcomes more often. As a result, it is challenging to disentangle whether an action was chosen based on value-based or value-free information.

Despite the success of RL in explaining human and animal behavior, recent work has shown that many empirical findings could also be partially explained by a value-free HL mechanism. In context-dependent preference learning, for example, humans may choose lower-value actions during transfer when those actions were paired with worse alternatives during training. While this phenomenon can be explained with RL assuming subjective evaluation of rewards (e.g., Bavard et al., 2021; Molinaro and Collins, 2023), a simple habit-like repetition bias can also account for it (Wagner et al., 2025). A similar repetition bias, irrespective of value, has also been found in a sequential decision-making task (Legler et al., 2025). Neuroscientific evidence further suggests that dopamine could signal not only reward prediction errors (as needed for model-free RL), but also action prediction errors, the teaching signal supporting value-free habit formation (Greenstreet et al., 2025).

Moreover, a recent proposal suggests that model-free RL-like behavior can be produced by combining HL with a working memory (WM) mechanism (Collins, 2025; see also Collins, 2018; Collins and Frank, 2012). WM is value-sensitive, but differs from model-free RL in two crucial ways: it learns associations almost immediately and can only retain a small set of associations for a short time. Thus, while WM can support adaptive behavior in the short run, HL can consolidate this adaptive behavior in the long run, yielding learning curves that are at first glance indistinguishable from RL. However, detailed error analysis showed that participants tended to repeat errors if WM was overloaded, a hallmark of HL, not RL. Correspondingly, a model combining HL and WM could explain error patterns better than models including model-free RL and/or WM[1]. This demonstrates that behavior can easily be misinterpreted as RL even if it was generated by a fundamentally different algorithm (Gu et al., 2025).

To directly investigate the separate contributions of HL and RL, we present two experiments ($n_{\text{total}} = 100$) that disentangle reward and choice histories while controlling for WM and subjective valuation. We achieved this by manipulating how often actions with different expected values were executed during training and measuring action preferences in a separate test phase. This allowed us to distinguish value-based and value-free learning components (accounting for WM): If only RL supports learning, then actions with higher expected value should be chosen during testing irrespective of how often they

[1]Note that HL differs from a choice perseveration kernel in RL models (e.g., Toyama et al., 2023) as it is stimulus-dependent and, in conjunction with other processes, an integral factor of learning.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

have been executed. Conversely, if only HL supports learning, then actions that have been executed more often should be chosen irrespective of their expected value.

We predicted that both RL and HL are involved in learning, resulting in the following preregistered hypotheses[2]:

1. **RL** — If the number of executions is equated during training, the probability of choosing the highest-value action is greater than that of the second-highest-value action. Furthermore, the probability of choosing the second-highest-value action increases the higher its value is.
2. **HL** — If the second-highest-value action is executed more often than the highest-value action during training, the probability of choosing the second-highest-value action during testing should either be higher than that of the highest-value action (**strong HL**) or be higher compared to the condition with equal executions (**moderate HL**).
3. The same effects should appear during training, when choosing between the highest-value and second-highest-value action.

We used computational models, including RL, HL, and their mixture to assess each component's contribution to behavior.

Methods

Experiment 1

Participants We preregistered a stopping rule of 50 participants after exclusions (41 female, 7 male, age 21.3 ± 3.7). All participants provided informed consent in accordance with the guidelines approved by the Ethics Committee at the University of California, Berkeley. The experiment terminated early if participants did not pass attention checks. We excluded 15 more participants because they did not meet the preregistered learning performance criterion of 70%. Participants were recruited through the research participation program for undergraduate students at the University of California, Berkeley and received course credit for their participation. We only allowed participants to take part if they were between the ages of 18 and 40, spoke English fluently, had no significant history of brain injury, mental/psychiatric illness like Parkinson's Disease, OCD, Schizophrenia, Depression, or ADD/ADHD, and no history of alcohol or drug abuse.

Task and Procedure Participants completed a computerized multi-armed bandit task consisting of at least 600 training trials, divided into eight blocks (two images per block; 16 images total; see top panel of Fig. 1). On each trial, participants saw an image and selected an action by pressing one of four keys. After each choice, participants received feedback of either "+1" or "0". The probabilities of receiving "+1" were [93%, 73%, 27%, 7%] for the best (A1), second-best (A2), third-best (A3), and worst action (A4). The action-key mapping for every block was chosen such that each combination occurred equally often throughout the experiment (Fig. 1B).

[2]Experiment 1: OSF preregistration. ; Experiment 2: OSF preregistration.

Each block was further divided into two sub-blocks, in each of which only two of the four actions were available. These were either the best and third-best actions (A1 and A3) or the second-best and worst actions (A2 and A4). This division ensured that the best and second-best actions were the better option in each sub-block with an equal difference in reward probability to the worse option. Participants transitioned from one sub-block to the next once they had reached a specified number of executions of either the best or second-best action (A1 and A2) for the two images shown in the block. In four of the eight blocks (showing eight of 16 images), A1 and A2 had to be selected 15 times each. In the remaining blocks (and images), A2 had to be selected 30 times and A1 15 times (Fig. 1B). Therefore, for half of the images, the second-best action was selected 15 times during training (*A2=A1*); for the other half of the images, it was selected 30 times (*A2>A1*). We did not control how often participants selected A3 and A4. However, the learning criterion of 70% ensured that these actions were on average selected less than 15 times. Additionally, the better and worse actions were always opposed between the two images in a sub-block. Participants were informed of this at the beginning of the experiment and instructed to collect as many rewards as possible.

After the training phase and a break of at least two minutes, participants completed a separate test phase. All 16 images were presented in random order with all four actions available and without feedback, to prevent further learning and WM use. The random sequence was repeated four times, ensuring maximal separation between the presentation of each image. Participants were not explicitly instructed to maximize rewards but to respond "as well as possible".

Design Experiment 1 used a within-subject design with two factorial independent variables: each action's reward probability (four levels) and the number of executions (two levels, A1=A2 or A2>A1, with eight of 16 images per level). Aggregating over images, we measured each action's choice frequency in the test phase. The assignment of four blocks to each condition, the blocks' order, and the assignment of images to blocks was randomized for each participant.

Experiment 2

Participants We had the same preregistered stopping rule of 50 participants after exclusions (22 female, 28 male, age 32.2 ± 4.9). All participants provided informed consent in accordance with the guidelines approved by the Ethics Committee at the University of California, Berkeley. The experiment terminated early if participants did not pass attention checks. We excluded 23 more participants because they did not meet the preregistered learning performance criterion of 60%. Participants were recruited through Prolific and received $12 as compensation. We used the same screening criteria as in Experiment 1.

Task and Procedure Experiment 2 was similar to Experiment 1, but presented three images in each of eight blocks

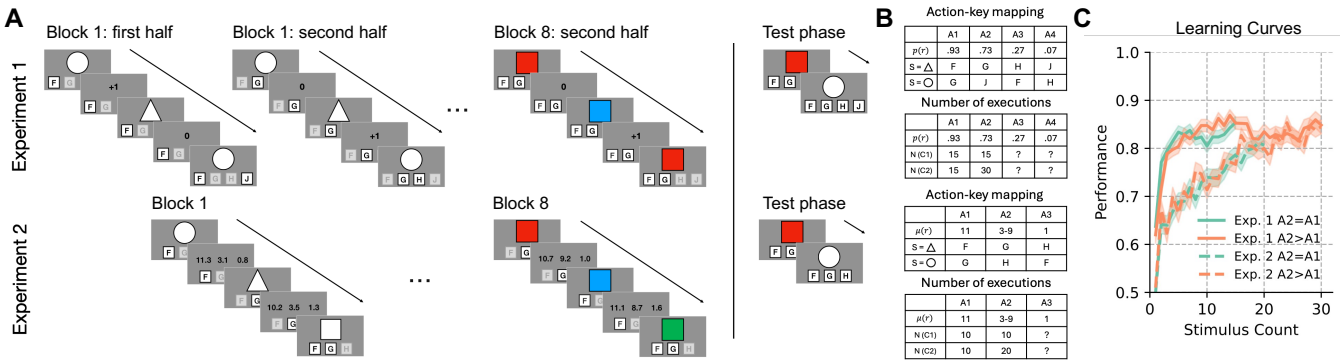

Figure 1: **Experimental Design and Learning Curves**. (**A**) Design schematic. Grayed-out keys denote unavailable actions (per block in Experiment 1 or trial in Experiment 2). Shape and color stimuli are placeholders for images used in the experiments. (**B**) Example action-key mappings; experimental condition manipulation of number of A2 executions. (**C**) Learning curves for the two experiments and conditions.

(24 images total) and used only three actions (Fig. 1, lower panel). Instead of manipulating reward probabilities per action, we used different reward magnitudes. The best action (A1) had an average magnitude of 11, the second-best action (A2) had an average magnitude of either 3, 5, 7, or 9, and the worst action (A3) had an average magnitude of 1. The exact reward was drawn on each trial from a normal distribution with a standard deviation of 0.3, such that the reward values virtually never overlapped between actions. The reward was always shown for all three actions, independently of whether they had been chosen or were available in the current trial.

At each trial, two of the three actions were available, indicated 500 ms before the stimulus was shown. Experiment 2 did not use separate sub-blocks as in Experiment 1, but the available actions changed at every trial (either A1-A3, A2-A3, or A1-A2). In the first condition (A2=A1, four of eight blocks; 12 of 24 images), A1 and A2 had to be chosen 10 times each to transition to the next block. In the second condition (A2>A1, remaining blocks and images), A2 had to be chosen 20 times and A1 10 times. This was achieved by presenting more trials in which only actions A2 and A3 were available. If A3 was chosen, this exact trial (with the same available actions) was repeated at a random time in the remaining trials. The order of available actions was pseudo-randomized such that the sets of available actions were distributed approximately evenly across each block.

The test phase was the same as in Experiment 1.

Design The experiment used a within-subject design with three factorial independent variables: each action's reward magnitude (three levels), the reward magnitudes for action A2 (four levels, with 6 of 24 images per level), and the number of executions of actions A1 and A2 (two levels, A1=A2 or A2>A1, with 12 of 24 images per level). Just as for Experiment 1, we measured the frequency with which each action was chosen in the test phase. Randomization was equivalent to Experiment 1.

Computational Modeling

The model implementation is based on Collins (2025). We did not implement the WM component, because we analyzed the test phase, where WM can be assumed to have little influence on the decision due to weight decay and high load.

RL Model Model-free RL was implemented as a standard delta-rule agent, tracking Q-values for each state-action pair with the update rule

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_{\text{RL}}[r_t - Q_t(s_t, a_t)]. \quad (1)$$

Q-values were initialized at $Q_0 = 1/n_A$, where n_A is the number of available actions in each training (sub-)block (two and three in Experiment 1 and 2, respectively). In Experiment 2, rewards were normalized by dividing by the maximal possible reward (12). The learning rate $\alpha_{\text{RL}} \in [0, 0.2]$ was a free parameter. We limited this range because of ceiling effects in high values (see *Model Fitting and Checks*).

The RL policy was computed by transforming Q-values through a standard softmax

$$\pi_{\text{RL}}(a|s) = \frac{\exp \beta Q(s, a)}{\sum_i \exp \beta Q(s, a_i)}. \quad (2)$$

We fixed $\beta = 25$ and added a noise component controlled through the free parameter $\epsilon \in [0, 1]$; this is a typical choice in this framework to improve identifiability (Collins, 2025; Senta et al., 2025), making the final policy

$$\pi(a|s) = (1 - \epsilon)\pi_{\text{RL}}(a|s) + \epsilon \frac{1}{n_A}. \quad (3)$$

HL Model HL was implemented with a slightly modified learning rule that did not include the reward r_t but simply tracked whether a state-action pair occurred

$$H_{t+1}(s_t, a_t) = H_t(s_t, a_t) + \alpha_{\text{HL}}[1 - H_t(s_t, a_t)]. \quad (4)$$

H-values were initialized in the same way as Q-values. The learning rate $\alpha_{\text{HL}} \in [0, 0.2]$ was free with the same range as for the RL component. The HL policy π_{HL} was computed in the same manner as the RL policy with $\beta = 25$ and $\epsilon \in [0, 1]$.

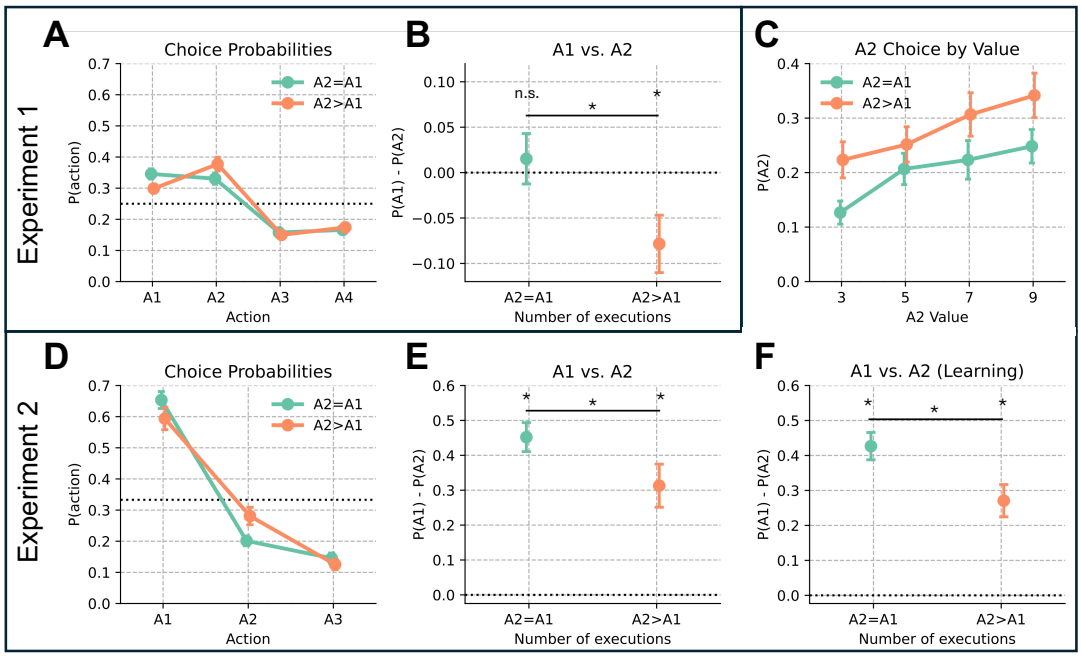

Figure 2: **Confirmatory Results.** The results support our preregistered hypotheses. A2=A1: Both actions were executed equally often during training; A2>A1: Action A2 was executed twice as often as action A1 during training. All results are from the test phase, except in panel F, where learning phase A1-A2 trials were used. Black dotted line indicates chance level. * indicates $p < .05$. Dots and error bars indicate means and SEMs.

RL+HL Model The RL+HL model used a weighted average of the RL and HL policies

$$\pi_{\text{mixture}}(a|s) = \omega\pi_{\text{RL}}(a|s) + (1-\omega)\pi_{\text{HL}}(a|s). \quad (5)$$

Parameter $\omega \in [0, 1]$ weighted the relative strength of value-based and value-free influences on participants' decisions.

Range Adaptation (RL-RA) Variant To explicitly test if our results could also be explained with range adaptation (Bavard et al., 2021), we implemented a variant of the RL model in which rewards are normalized based on the range of rewards in the current decision context

$$r_{\text{range}} = \frac{r - r_{\text{min,t}}(c)}{r_{\text{max,t}}(c) - r_{\text{min,t}}(c)}, \quad (6)$$

where $r_{\text{min,t}}(c)$ is the minimal reward in decision context c encountered up to the current trial t and $r_{\text{max,t}}(c)$ is the corresponding maximal reward. At $t = 0$, both reward bounds were initialized at zero and updated at every trial. The adapted reward was only computed once $r_{\text{max,t}}(c) \neq 0$ to avoid division by zero.

Model Fitting and Checks We fit parameters to each participant's test phase data individually, conditioned on model values derived through their own choice and reward history in the learning phase (two parameters for RL, RL-RA, and HL, and four parameters for the RL+HL model) using maximum likelihood estimation with SciPy (Virtanen et al., 2020). We did not fit to the training phase because the current design did not enable us to identify the influence of WM during training.

Parameter recoverability analyses (Wilson & Collins, 2019) suggested that the weighting parameter ω and the noise parameter ϵ were adequately identifiable (Spearman's $\rho(\omega) = (.57, .73)$ and $\rho(\epsilon) = (.79, .78)$, for the first and second experiments, respectively). Because we did not fit to the learning phase, the learning rate parameters were poorly identifiable (Spearman's $\rho(\alpha_{\text{RL}}) = (.44, .01)$ and $\rho(\alpha_{\text{HL}}) = (.01, .03)$). We therefore do not report the fitted values of these two parameters and do not draw any conclusions based on them (Wilson & Collins, 2019).

Model recoverability analyses demonstrated that the RL/RL-RA, HL, and the RL+HL models were highly identifiable based on AIC (probability of correct recovery was at least 97% in both experiments). The RL-RA could not be distinguished well from the standard RL model (probability of correct recovery 44-56%) because the experiment was specifically designed to differentiate range adaptation from HL not RL. Identifiability based on BIC was worse for the RL+HL model. For posterior predictive checks, we simulated model behavior in the test phase, conditioned on the model variables learned through each participant's experienced image-action(-reward) combinations in the learning phase.

Results

Confirmatory Results

We verified that participants learned as expected in both experiments' learning phases (Fig. 1C). Participants' choices in the test phase (Fig. 2A and D) confirmed that they retained

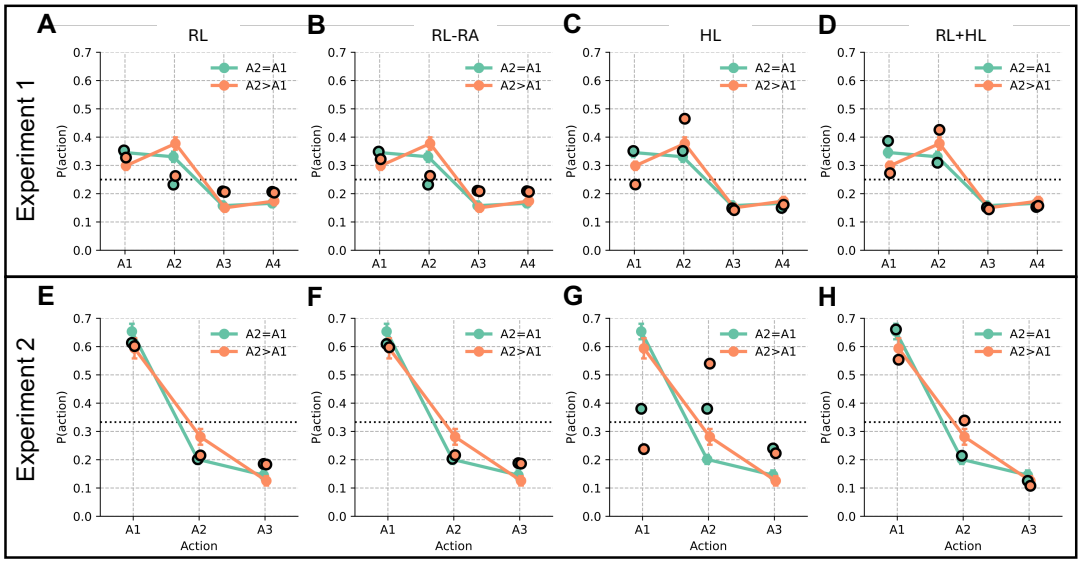

Figure 3: **Model predictions**. Behavioral data (lines; A-D: Experiment 1, E-H: Experiment 2) and model predictions (black circles; fill color indicates condition). A2=A1: Both actions were executed equally often during training; A2>A1: Action A2 was executed twice as often as action A1 during training. Only the RL+HL model captured the behavioral effects.

learning phase information. As preregistered, we computed the difference in choice probabilities between the best and second-best actions (A1 and A2) for each condition (Fig. 2B and E). Here, values above zero mean that action A1 was chosen more often than A2. In Experiment 1, participants did not choose A1 over A2 if they were executed equally often, $t(49) = 0.55, p = .587$, but they chose A2 over A1 if A2 was executed more often than A1, $t(49) = -2.48, p = .017, d = -0.35$ (Fig. 2B). Correspondingly, the tendency to choose A2 over A1 increased from the equal to the unequal condition, $t(49) = -2.14, p = .038, d = -0.30$. This suggests that participants' choices were influenced primarily by their choice history, not their reward history, and provides evidence in favor of a strong HL hypothesis, as the influence of choice seems to have overruled any potential influence of reward.

In Experiment 2, participants chose A1 over A2 if both actions were selected equally often, $t(49) = 10.88, p < .001, d = 1.54$, and still did so if A2 was selected twice as often as A1, $t(49) = 5.06, p < .001, d = 0.72$ (Fig. 2E). However, the difference in choice probabilities increased in favor of action A2 if this action was selected more often, $W = 377.5, p = .012, r = .34$. As preregistered, we deployed a Wilcoxon signed-rank test as the difference scores could not be assumed to be approximately normal. The same effects were present during the training phase, considering only trials in which A1 and A2 were directly compared (Fig. 2F). A1 was chosen over A2 in both conditions, $t(49) = 10.93, p < .001, d = 1.55$, $t(49) = 5.87, p < .001, d = 0.83$, and the difference in choice probability increased, $W = 169.0, p < .001, r = .30$. Thus in Experiment 2, participants' choices seemed to be driven primarily by their reward history with an additional but weaker influence of their choice history. A mixed-effect logistic regression revealed that the probability of choosing A2 increased with A2's average value $OR = 1.12, 95\%CI = [1.06, 1.17], p < .001$ (Fig. 2C). This further corroborates that participants were sensitive to the action values. Yet, even for high values of A2, the choice probability was below that of A1, indicating clear sensitivity to the difference between average rewards of 9 and 11.

Modeling Results

Figure 3A-H shows posterior predictions for test phase behavior for the four models. In Experiment 1, the HL model qualitatively reproduces the patterns of results (Fig. 3C), whereas the RL and RL-RA models cannot (Fig. 3A-B). The RL+HL model provides a closer quantitative fit (Fig. 3D). In Experiment 2, the RL and RL-RA models predict participants' reward sensitivity (choice probability of A1 higher than of A2 and A3; Fig. 3E-F) but, crucially, cannot explain the effect of A2 executions frequency during training (increase in choice probability of A2 at test). Only the RL+HL model predicts this effect accurately (Fig. 3H).

As Table 1 shows, the HL model performed best in Experiment 1. Correspondingly, the RL+HL model put a higher weight on the HL component for the majority of participants (Fig. 4A). Only a small number of participants received a high weight for the RL component, indicating that most participants exhibited a value-independent response behavior. In Experiment 2, the HL model performed worse, with the RL+HL model showing the overall best performance. In line with this, the RL component was weighted more strongly for the majority of participants (Fig. 4B). Overall, the modeling analyses corroborate the behavioral results, demonstrating a mixture of value-free and value-based influences on responses. The

	Experiment 1	Experiment 2
RL	8571.29	8123.68
RL-RA	8587.82	8148.75
HL	**8146.76**	10239.06
RL+HL	8349.82	**7778.41**

Table 1: **Model comparison via AIC**

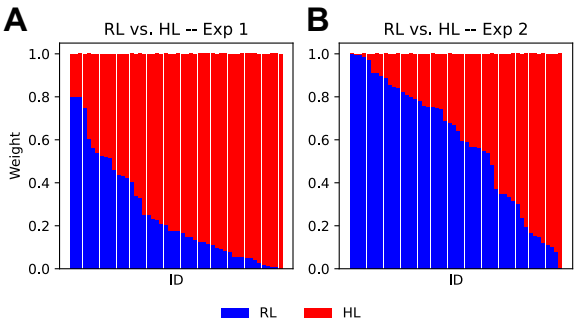

Figure 4: **Mixture weight of RL and HL in the RL+HL Model**. A purely blue (red) vertical line indicates a weight of 1 for the RL (HL) component for that participant. (**A**): Experiment 1. (**B**) Experiment 2.

composition of this mixture varies between mostly value-free (in Experiment 1) and mostly value-based (in Experiment 2). This suggests that whether behavior is dominated by reward or choice history depends on environmental circumstances. Range adaptation cannot provide an alternative explanation for the observed value-free learning effect.

Exploratory Results

To further assess the results, we fit a hierarchical Bayesian multinomial logistic regression model predicting action choice (with A1 as reference action) based on the number of A2 executions during training (and the value of A2 in Experiment 2; using random intercepts for participants). In Experiment 1, the probability of choosing A2 over A1 increased during testing if A2 was chosen more often during training, $\beta = 0.28, \text{SD} = 0.09, 99\%\ \text{HDI} = [0.07, 0.51]$. In Experiment 2, the probability of choosing A2 over A1 also increased, $\beta = 0.44, \text{SD} = 0.13, 99\%\ \text{HDI} = [0.12, 0.78]$. Furthermore, the probability of choosing A2 over A1 during testing increased as the value of A2 during training increased, $\beta = 0.64, \text{SD} = 0.15, 99\%\ \text{HDI} = [0.28, 1.02]$. This supports the confirmatory results, suggesting a value-independent influence of choice frequency on test behavior.

Discussion

This study investigated to what extent human behavior is driven by the choices made in the past or the consequences of those choices. We disentangled these two factors experimentally while controlling for other mechanisms such as working memory (WM, by focusing only on the test phase) and range adaptation (by equating the range). Our results provide evidence in favor of a combination of choice-based and outcome-based influences on behavior. Which influence dominates seems to depend on characteristics of the task environment.

Our results indicate that without a carefully controlled experimental design, behavior can be misinterpreted as being solely driven by model-free reinforcement learning (RL, i.e., outcome-based learning) even if it is not. This is in line with research suggesting alternative learning mechanisms, such as WM (Collins, 2018) or episodic memory (Bornstein and Norman, 2017). Recent work has also specifically drawn a distinction between habitual learning (HL) and model-free RL both theoretically (Miller et al., 2019) and empirically (Collins, 2025; Greenstreet et al., 2025; Legler et al., 2025; Wagner et al., 2025). By explicitly manipulating choice and reward history separately, we provide direct evidence that HL contributes to learning, and that overlooking this contribution may lead researchers to mis-identify the neurocognitive processes that underlie choices (see also O'Doherty, 2014).

The relative influence of HL and RL on behavior varied considerably between the two experiments. One reason we found little evidence of RL in Experiment 1 may be that test phase performance was noisy, limiting our power to detect RL. In the learning phase, participants quickly achieved their asymptotic level of performance, suggesting that learning might have relied strongly on WM; recent findings show that WM might attenuate RL if WM load is low (Collins, 2018). Experiment 2 imposed a higher WM load (more simultaneous stimuli and actions), possibly explaining a stronger influence of RL. While RL seems to gain influence in the absence of WM, HL crucially relies on WM to produce adaptive behavior (see Collins, 2025). This suggests that WM might play a fundamental role in the interplay of RL and HL, posing an important avenue for future research.

Furthermore, while Experiment 1 used partial reward feedback (only for the chosen action) and binary rewards with different reward probabilities, Experiment 2 showed reward feedback for all actions and used non-overlapping reward magnitudes. One could therefore argue that reward feedback was more informative and directly available in Experiment 2, where we also found stronger evidence of RL. This is consistent with the notion that humans might rationally rely on both value-free and value-based information, depending on their usefulness (Miller et al., 2019). Previous modeling work also showed that augmenting a model-free RL algorithm with an HL component can yield more stable and adaptive performance (Greenstreet et al., 2025), justifying the use of value-free information beyond a mere reduction in complexity (e.g., Gershman, 2020).

In summary, we provide evidence of separable influences of model-free RL, HL, and WM on behavior. It is easy to misinterpret these learning mechanisms as a single process, such as model-free RL. However, this misinterpretation could strongly affect the conclusions we draw about human behavior and their real-world implications.

Acknowledgments

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and the German Academic Exchange Service (DAAD; research stipend for doctoral students, 2024) for supporting Maximilian Mittenbühler. This project was in part supported by NIMH R21H136528 to AGEC.

We acknowledge the use of large language model (LLM) tools for supporting language checks and smaller coding tasks, such as data visualization.

References

Bavard, S., Rustichini, A., & Palminteri, S. (2021). Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning. *Science Advances*, *7*(14), eabe0340. https://doi.org/10.1126/sciadv.abe0340

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, *20*(7), 997–1003. https://doi.org/10.1038/nn.4573

Collins, A. G. E. (2018). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432. https://doi.org/10.1162/jocn_a_01238

Collins, A. G. E. (2025). A habit and working memory model as an alternative account of human reward-based learning. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-025-02340-0

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1036–1051. https://doi.org/10.1111/j.1460-9568.2012.08050.x

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, *204*, 104394. https://doi.org/10.1016/j.cognition.2020.104394

Greenstreet, F., Vergara, H. M., Johansson, Y., Pati, S., Schwarz, L., Lenzi, S. C., Geerts, J. P., Wisdom, M., Gubanova, A., Rollik, L. B., Kaur, J., Moskovitz, T., Cohen, J., Thompson, E., Margrie, T. W., Clopath, C., & Stephenson-Jones, M. (2025). Dopaminergic action prediction errors serve as a value-free teaching signal. *Nature*, *643*(8074), 1333–1342. https://doi.org/10.1038/s41586-025-09008-9

Gu, X., Jiang, C., & Shi, L. (2025, November). Mimicking Human Intuition: Cognitive Belief-Driven Reinforcement Learning [arXiv:2410.01739 [cs]]. https://doi.org/10.48550/arXiv.2410.01739

Legler, E., Cuevas Rivera, D., Schwöbel, S., Wagner, B. J., & Kiebel, S. (2025). Cognitive computational model reveals repetition bias in a sequential decision-making task. *Communications Psychology*, *3*(1), 92. https://doi.org/10.1038/s44271-025-00271-0

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, *126*(2), 292–311. https://doi.org/10.1037/rev0000120

Molinaro, G., & Collins, A. G. E. (2023). Intrinsic rewards explain context-sensitive valuation in reinforcement learning (C. Summerfield, Ed.). *PLOS Biology*, *21*(7), e3002201. https://doi.org/10.1371/journal.pbio.3002201

O'Doherty, J. P. (2014). The problem with value. *Neuroscience & Biobehavioral Reviews*, *43*, 259–268. https://doi.org/10.1016/j.neubiorev.2014.03.027

Senta, J. D., Bishop, S. J., & Collins, A. G. (2025). Dual process impairments in reinforcement learning and working memory systems underlie learning deficits in physiological anxiety (L. J. Graham, Ed.). *PLOS Computational Biology*, *21*(9), e1012872. https://doi.org/10.1371/journal.pcbi.1012872

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

Toyama, A., Katahira, K., & Kunisato, Y. (2023). Examinations of Biases by Model Misspecification and Parameter Reliability of Reinforcement Learning Models. *Computational Brain & Behavior*, *6*(4), 651–670. https://doi.org/10.1007/s42113-023-00175-4

Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wagner, B. J., Wolf, H. B., & Kiebel, S. J. (2025). Action repetition biases choice in context-dependent decision-making. *Communications Psychology*, *3*(1), 177. https://doi.org/10.1038/s44271-025-00363-x

Watson, P., O'Callaghan, C., Perkes, I., Bradfield, L., & Turner, K. (2022). Making habits measurable beyond what they are not: A focus on associative dual-process models. *Neuroscience & Biobehavioral Reviews*, *142*, 104869. https://doi.org/10.1016/j.neubiorev.2022.104869

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Wood, W., Mazar, A., & Neal, D. T. (2022). Habits and Goals in Human Behavior: Separate but Interacting Systems. *Perspectives on Psychological Science*, *17*(2), 590–605. https://doi.org/10.1177/1745691621994226

Wood, W., & Rünger, D. (2016). Psychology of Habit. *Annual Review of Psychology*, *67*(1), 289–314. https://doi.org/10.1146/annurev-psych-122414-033417