PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Active reinforcement learning versus action bias and hysteresis: control with a mixture of experts and nonexperts

Jaron T. Colas[1,2,3]*, **John P. O'Doherty**[2,3‡], **Scott T. Grafton**[1‡]

1 Department of Psychological and Brain Sciences, University of California, Santa Barbara, California, United States of America, **2** Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California, United States of America, **3** Computation and Neural Systems Program, California Institute of Technology, Pasadena, California, United States of America

‡ These authors are joint senior authors on this work.
* jcolas@ucsb.edu

Citation: Colas JT, O'Doherty JP, Grafton ST (2024) Active reinforcement learning versus action bias and hysteresis: control with a mixture of experts and nonexperts. PLoS Comput Biol 20(3): e1011950. https://doi.org/10.1371/journal.pcbi.1011950

Editor: Stefano Palminteri, Ecole Normale Superieure, FRANCE

Received: September 13, 2023

Accepted: February 26, 2024

Published: March 29, 2024

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: STG was supported by the Institute for Collaborative Biotechnologies under Cooperative Agreement W911NF-19-2-0026 and grant W911NF-16-1-0474 from the Army Research Office. JPOD was supported by National Institute on Drug Abuse grant R01 DA040011 and the National Institute of Mental Health's Caltech Conte Center for Social Decision Making (P50

Abstract

Active reinforcement learning enables dynamic prediction and control, where one should not only maximize rewards but also minimize costs such as of inference, decisions, actions, and time. For an embodied agent such as a human, decisions are also shaped by physical aspects of actions. Beyond the effects of reward outcomes on learning processes, to what extent can modeling of behavior in a reinforcement-learning task be complicated by other sources of variance in sequential action choices? What of the effects of action bias (for actions per se) and action hysteresis determined by the history of actions chosen previously? The present study addressed these questions with incremental assembly of models for the sequential choice data from a task with hierarchical structure for additional complexity in learning. With systematic comparison and falsification of computational models, human choices were tested for signatures of parallel modules representing not only an enhanced form of generalized reinforcement learning but also action bias and hysteresis. We found evidence for substantial differences in bias and hysteresis across participants—even comparable in magnitude to the individual differences in learning. Individuals who did not learn well revealed the greatest biases, but those who did learn accurately were also significantly biased. The direction of hysteresis varied among individuals as repetition or, more commonly, alternation biases persisting from multiple previous actions. Considering that these actions were button presses with trivial motor demands, the idiosyncratic forces biasing sequences of action choices were robust enough to suggest ubiquity across individuals and across tasks requiring various actions. In light of how bias and hysteresis function as a heuristic for efficient control that adapts to uncertainty or low motivation by minimizing the cost of effort, these phenomena broaden the consilient theory of a mixture of experts to encompass a mixture of expert and nonexpert controllers of behavior.

MH094258). The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Reinforcement learning unifies neuroscience and AI with a universal computational framework for motivated behavior. Humans and robots alike are active and embodied agents who physically interact with the world and learn from feedback to guide future actions while weighing costs of time and energy. Initially, the modeling here attempted to identify learning algorithms for an interactive environment structured with patterns in counterfactual information that a human brain could learn to generalize. However, behavioral analysis revealed that a wider scope was necessary to identify individual differences in not only complex learning but also action bias and hysteresis. Sequential choices in the pursuit of rewards were clearly influenced by endogenous action preferences and persistent bias effects from action history causing repetition or alternation of previous actions. By modeling a modular brain as a mixture of expert and nonexpert systems for behavioral control, a distinct profile could be characterized for each individual attempting the experiment. Even for actions as simple as button pressing, effects specific to actions were as substantial as the effects from reward outcomes that decisions were supposed to follow from. Bias and hysteresis are concluded to be ubiquitous and intertwined with processes of active reinforcement learning for efficiency in behavior.

Introduction

Whether in machine learning and artificial intelligence or in animal learning and neural intelligence, the most crucial portion of reinforcement learning (RL) [1–3] is not passive, offline, or observational but instead active and online with a challenge of not only prediction but also real-time control. In the real world, resources for activity are finite, and much of active RL is also embodied RL. Whether robot or human, the embodied agent learns from feedback to make decisions and select physical actions that maximize future reward while minimizing various costs of energy as well as time.

The RL framework has appreciable predictive validity [4,5] when accounting for human choices and learning behavior in a variety of settings [6–8]—let alone the power of extensions of RL [9–12]. However, such models sometimes fail to account well for an individual's behavior even in a relatively simple task that should be amenable to RL in principle [13]. An open question concerns whether other components of variance not based on learning also exist alongside RL so as to collectively provide a better account of motivated behavior and even learning itself within a more comprehensive model. The present study focuses on the contributions of other elements of active learning that are also essential in their own way: action bias—specifically for actions per se—and action hysteresis, which is determined by the history of previously selected actions (**Fig 1A**).

The present case of two available actions (one per hand) reduces the first component of action bias to a single bidirectional constant for left versus right [14–16]. Hysteresis is bidirectional as well and adds dynamics in the form of either repetition or alternation of previous actions, which may also manifest for a horizon beyond just the most recent action [17–20]. Despite at least some precedent for either action bias or action hysteresis (more so the latter), the combination of both bias and hysteresis has even less precedent for RL [12,21].

The standard setup for fitting RL to behavior (e.g., [22]) begins with a 2-parameter model tuned for the learning rate and the softmax temperature, where the latter represents stochasticity [3,23–25]. This base model is then built upon with additional free parameters to test for more complex learning phenomena, which should include the due diligence of model

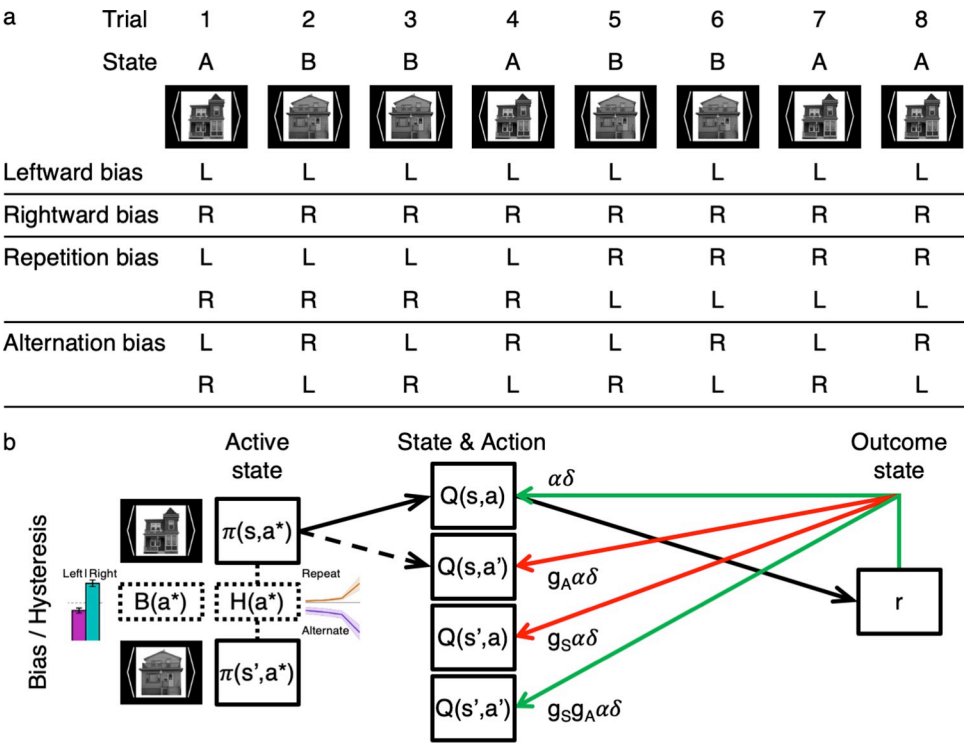

Fig 1. Action bias and hysteresis for the "generalized reinforcement learning" (GRL) model. (a) Each trial of the structured reward-learning task was initiated with an image cue symbolizing the state of the environment (e.g., "A" or "B"), where the optimal action given the state was a button press with either the left ("L") or right ("R") hand. In contrast to the expert control of GRL for mapping state-action pairs to rewards, the nonexpert forces of action bias and hysteresis were modeled as leftward or rightward bias and repetition or alternation bias. These action-specific effects manifest independently of the external state and reward history. **(b)** What matters for the present purposes is that, while a model with GRL adds complexity to basic RL, even more complexity must be accommodated for action bias and hysteresis. The agent's mixture policy $\pi_t(s_t,a^*)$ is probabilistic over available actions a^* in state s_t. The action selection of this mixture policy is determined by not only learned value for state-action pairs $Q_t(s_t,a^*)$ but also constant bias $B(a^*)$ and dynamic hysteretic bias $H_t(a^*)$ with an exponentially decaying hysteresis trace. The outcome of the chosen action a_t is a reward r_{t+1} that updates $Q_t(s_t,a_t)$ via the reward-prediction error (RPE) δ_{t+1} weighted by a learning rate α. For GRL specifically, this RPE signal is generalized to representations of other state-action pairs according to extra parameters for action generalization (g_A) and state generalization (g_S). See Figs 8 and 13 for details of the plots representing individual differences in constant lateral bias (left versus right) and the exponential hysteresis trace (repeat versus alternate). See also the original report of this study with additional details about the paradigm and GRL per se [12].

https://doi.org/10.1371/journal.pcbi.1011950.g001

comparison and qualitative falsification [26–28]. However, an alternative line of questioning could instead begin with asking whether more parsimonious and perhaps more substantial sources of variance merit prioritization before making any new assumptions about complexities within learning. The emphasis can also be shifted away from the prescriptive (i.e., "According to some notion of 'optimality', what should a person do here?") in favor of the descriptive ("What do people actually do here?") while creating an opportunity to circle back from empirical findings to a new perspective on different aspects of optimality in behavior.

In practice, model fitting is nontrivial with a sequence of choices typically limited to hundreds or even just dozens of observations. Adding to this challenge, increasingly complex behavior under study imposes greater demands for accommodating multidimensional individual differences and optimizing individual fits without hierarchical Bayesian fitting [13,29] and its disadvantage of estimation bias [30–35]. (For a random grouping of independent data sets, even hierarchical fitting compromises their independence with the strong assumption of a

common distribution for every individual based on the ecological fallacy [36–38].) Both within and between individual sequences, sources of variance other than RL may be crucial to complement an RL model despite the costs of additional degrees of freedom. In other words, including modules beyond RL in a model of actual behavior can alleviate estimation bias and other distortions of learning parameters that would otherwise be forced to simultaneously fit other phenomena with omitted variables.

In the present study, we hypothesize that behavior during active learning is determined not only by RL and stochasticity but also by action bias and hysteresis, which are independent of the current state of the external environment and its reward history (**Fig 1**). This state-independent hysteresis in particular makes actions depend on previous actions regardless of states, but state-dependent hysteresis was also considered later (**Table 1**). The interplay of these different forces was investigated for human behavior in a task that in one sense is a hierarchical reversal-learning task but in another sense is a sequential button-pressing task (**Fig A in S1 Text**). Hence the behavioral data of a multisite neuroimaging study reported previously [12] were reanalyzed with further model comparison from this bias-centric perspective.

Too often, such action-specific effects have been overlooked altogether or given only cursory mention as if they were inconsequential in the context of a learning model. If considered at all, the scope of hysteresis has also usually been limited to only one trial back. (To address this issue here, we modeled hysteresis over a time horizon longer than one trial.) Moreover, because repetition tends to predominate in aggregate behavior for RL and other sequential paradigms, manifestations of hysteresis have mostly been framed so as to deemphasize or entirely disregard alternation biases in favor of repetition biases. Autocorrelational effects have thus been referred to in the literature with unidirectional and often imprecise terminology such as "perseveration", "perseverance" (a misnomer), "persistence", "habit", "choice stickiness", "choice consistency", "repetition priming", "response inertia", or "behavioral momentum". Semantics of interpretation aside, the common thread for hysteresis is a past action's influence on an upcoming action with independence from learnable external feedback and typically, albeit not necessarily, from external states as well.

A more comprehensive model of action selection can also enhance identifiability with respect to actual learning (or lack thereof) as opposed to other components of variance that may mimic or otherwise obscure signatures of learning with spurious correlations across the finite sequence of actions [17,18,27,28,39–47]. As external reinforcement promotes consistent repetition of responses within a state, so too can action bias, and both repetition and alternation from hysteresis can coincidentally align with the reward contingencies of the sequence of states. Whereas preexisting constant biases interact with learning when base rates for actions are unbalanced in sequence, hysteretic biases can further complicate action sequences with not only intrinsic dynamics but also more possibilities for interactions across any sequential patterns in the environment and the dynamics of learning.

Perhaps surprisingly, the hypothesis for hysteresis in the present experiment was that alternation would predominate rather than repetition. An action policy biased toward alternation would follow from the fact that, by design, choosing actions optimally in response to the rotating states of this environment would result in alternating more frequently. Yet, by design, this perseverative alternation that is characteristically independent of learned external value was therefore not conducive to obtaining more rewards from this environment.

The primary model comparison here (**Table 2 and Table A in S1 Text**) exhaustively tested various combinations of action-specific effects as well as "generalized reinforcement learning" (GRL), which is a quasi-model-based extension of model-free RL that can flexibly generalize value information across states and actions (**Fig 1B and Fig B in S1 Text**) [12]. GRL per se is somewhat incidental for the present purposes, but what matters as far as a test case here is that

Table 1. Variables for basic forms of RL, bias, and hysteresis. Fundamentally for even basic RL, the possibilities for variables in a more comprehensive behavioral model can be classified according to dependence on (or independence of) states, actions. previous actions, and reward outcomes. In principle, whereas action value is outcome-dependent, action hysteresis is outcome-independent. However, when modeling actual behavior, this conceptual independence does not guarantee statistical independence because of incidental correlations in finite sequences of action choices. For the present study, the primary model comparison focuses on the three variables (marked with an asterisk) that are the most fundamental and typically the most dissociable—namely, constant bias *B(a)*, state-independent action hysteresis *H(a)*, and state-dependent action value *Q(s,a)*. The extended model comparison also incorporates state-dependent action hysteresis *H(s,a)* and state-independent action value *Q(a)*. Note that state value *V(s)* is generally relevant in RL but is not considered here. The abbreviations "PrevAction", "dep.", and "indep." correspond to "previous action", "dependent", and "independent", respectively.

Variable	Term	Action-	PrevAction-	State-	Outcome-
Constant action bias*	B(a)	dep.	indep.	indep.	indep.
State-independent action hysteresis*	H(a)	dep.	dep.	indep.	indep.
State-dependent action hysteresis	H(s,a)	dep.	dep.	dep.	indep.
State-independent action value	Q(a)	dep.	indep.	indep.	dep.
State-dependent action value*	Q(s,a)	dep.	indep.	dep.	dep.
State value	V(s)	indep.	indep.	dep.	dep.

https://doi.org/10.1371/journal.pcbi.1011950.t001

a model incorporating the complexities of bias and hysteresis should still be amenable to exploring complex learning algorithms beyond the most basic RL. GRL is especially complicating in this regard because it introduces high-frequency dynamics to learning with counterfactual updates of multiple value representations in parallel.

Previously, the GRL model was built with fixed prior assumptions for another three free parameters representing action bias and hysteresis. One of these parameters specifies the constant lateral bias; the other two specify a decaying exponential function for the hysteresis trace extending backward across the sequence. This particular configuration of constant bias and exponential hysteresis was initially arrived at intuitively more so than empirically [12,21] while drawing elements from earlier models [17,18]. Now, the 3-parameter adjunct was to actually be tested against GRL alone as well as both simpler and more complex variations for bias and (state-independent) hysteresis. Subsequent testing also proceeded to alternative model features that could be other sources of action repetition or alternation, including state-dependent hysteresis, state-independent action value, confirmation bias in learning, or asymmetric learning rates more generally.

Abiding by Occam's razor [48], the more parsimonious factors of action bias and hysteresis should be granted first priority for inclusion if they are sufficiently substantial, but testing empirical data was necessary to verify practical feasibility in consideration of the compounded complexity with different forms of learning. Individuals found to not learn well were expected to reveal the greatest effects of bias and hysteresis. Yet those who learned accurately were also hypothesized to exhibit biases that would account for significant variance (even if this were to amount to less variance than that from learning).

To the end of establishing guidelines for behavioral modeling in general, there were further questions concerning how exactly these directional biases would manifest and how substantial they would be for the experimenter's default choice of pressing a button, which is a simple and familiar action with trivial motor demands. For proof of concept, the present paradigm can query not only the suitability of these particular forms of biases for button presses but also the viability of these factors as additional complexities while learning theory is advanced. With reference to analogous architectures in machine learning [49–54] as well as with general appeal to modular parallelism and conditional computation for balancing versatility and efficiency in optimal control, the consilient theory of a mixture of experts [6–8,55–57] can be broadened further for a mixture of expert and nonexpert controllers of behavior (see Discussion). This contrast of expertise versus efficiency is represented here by different types of expert RL versus nonexpert bias and hysteresis.

Table 2. Model parameters (condensed). Free parameters are listed for the 72 behavioral models in ascending order of complexity within and across classes. The models are coded with the first letter of the label referring to four possibilities: an absence of learning ("X"), reinforcement learning (RL) without generalization ("0"), generalized reinforcement learning (GRL) with one shared generalization parameter g_1 ("1"), or GRL with two separate generalization parameters g_1 and g_2 ("2"). RL itself required free parameters for the learning rate α and the softmax temperature τ. Models labeled with "C" for the second letter included a constant lateral bias, which was arbitrarily designated as a rightward bias β_R (where $\beta_R < 0$ is leftward). The list is condensed with bracket notation to represent the range for the n-back horizons of each successive model within a hysteresis category (e.g., "2CE[1–3]" for models 2CE1, 2CE2, and 2CE3). Models labeled with"N" and ending with a positive integer (from the range in brackets) included n-back hysteresis with free parameters β_n for repetition ($\beta_n > 0$) or alternation ($\beta_n < 0$) of each previous action represented—up to 4 trials back (β_4) with learning and up to 8 trials back (β_8) without learning. Models labeled with "E" and ending with a positive integer N (from the range in brackets) included exponential hysteresis with inverse decay rate λ_H taking effect $N+1$ trials back. Exponential models could also be both parametric and nonparametric with N free parameters β_n for initial n-back hysteresis up to 3 trials back (β_3), where the final β_N is the initial magnitude of the exponential component. "df" stands for degrees of freedom. See also Table A in S1 Text for the unrolled version of the list. This ordering of the models corresponds to the ordering in Figs 2 and 3.

		RL		GRL		Bias	Hysteresis								
Model	**df**	α	τ	g_1	g_2	β_R	λ_H	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
X	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XC	1	-	-	-	-	β_R	-	-	-	-	-	-	-	-	-
XN[1–8]	1–8	-	-	-	-	-	-	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
XCN[1–8]	2–9	-	-	-	-	β_R	-	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
XE[1–3]	2–4	-	-	-	-	-	λ_H	β_1	β_2	β_3	-	-	-	-	-
XCE[1–3]	3–5	-	-	-	-	β_R	λ_H	β_1	β_2	β_3	-	-	-	-	-
0	2	α	τ	-	-	-	-	-	-	-	-	-	-	-	-
0C	3	α	τ	-	-	β_R	-	-	-	-	-	-	-	-	-
0N[1–4]	3–6	α	τ	-	-	-	-	β_1	β_2	β_3	β_4	-	-	-	-
0CN[1–4]	4–7	α	τ	-	-	β_R	-	β_1	β_2	β_3	β_4	-	-	-	-
0E[1–3]	4–6	α	τ	-	-	-	λ_H	β_1	β_2	β_3	-	-	-	-	-
0CE[1–3]	5–7	α	τ	-	-	β_R	λ_H	β_1	β_2	β_3	-	-	-	-	-
1	3	α	τ	g_1	-	-	-	-	-	-	-	-	-	-	-
1C	4	α	τ	g_1	-	β_R	-	-	-	-	-	-	-	-	-
1N[1–4]	4–7	α	τ	g_1	-	-	-	β_1	β_2	β_3	β_4	-	-	-	-
1CN[1–4]	5–8	α	τ	g_1	-	β_R	-	β_1	β_2	β_3	β_4	-	-	-	-
1E[1–3]	5–7	α	τ	g_1	-	-	λ_H	β_1	β_2	β_3	-	-	-	-	-
1CE[1–3]	6–8	α	τ	g_1	-	β_R	λ_H	β_1	β_2	β_3	-	-	-	-	-
2	4	α	τ	g_1	g_2	-	-	-	-	-	-	-	-	-	-
2C	5	α	τ	g_1	g_2	β_R	-	-	-	-	-	-	-	-	-
2N[1–4]	5–8	α	τ	g_1	g_2	-	-	β_1	β_2	β_3	β_4	-	-	-	-
2CN[1–4]	6–9	α	τ	g_1	g_2	β_R	-	β_1	β_2	β_3	β_4	-	-	-	-
2E[1–3]	6–8	α	τ	g_1	g_2	-	λ_H	β_1	β_2	β_3	-	-	-	-	-
2CE[1–3]	7–9	α	τ	g_1	g_2	β_R	λ_H	β_1	β_2	β_3	-	-	-	-	-

https://doi.org/10.1371/journal.pcbi.1011950.t002

Results

Paradigm

Additional details of the study and previous results can be found in the original report for these data sets [12]. The hierarchical reversal-learning task delivered probabilistic outcomes for combinations of categorized states and contingent actions with reward distributions changing across 12 blocks of trials (**Figs A and B in S1 Text**). Suitably for first testing GRL, the state (or context) of each trial represented a two-armed contextual bandit belonging to one of two categories (e.g., faces or houses) with two anticorrelated states per category and two anticorrelated actions per state (i.e., left-hand button press or right-hand button press). For an optimal learner, the counterfactual information in this anticorrelational structure could be leveraged with the discriminative generalization of GRL. The action-generalization weight g_A

and state-generalization weight g_S, which would ideally both be negative for discriminative generalization, govern the relaying of the reward-prediction error across state-dependent actions or across states within a category, respectively.

For standard behavioral RL (with or without an extension such as GRL), the state-dependent action values $Q_t(s,a)$ that are learned over time would be the only inputs to a probabilistic action-selection policy $\pi_t(s,a)$ characterized by a softmax function with temperature τ:

$$\pi_t(s_t, a) = P(a_t = a|s_t) = \frac{\exp\{Q_t(s_t, a)/\tau\}}{\sum_{a^*}\exp\{Q_t(s_t, a^*)/\tau\}}$$

As the scope of the model is expanded, the present study emphasizes that the action policy is a function of not only action value $Q_t(s,a)$ but also constant bias $B(a)$ and dynamic hysteretic bias $H_t(a)$ as modules within a mixture of experts and nonexperts (**Fig 1**) [12,21]. Constant bias $B(a)$ becomes a lateral bias between left and right actions in this case, whereas the dynamic hysteretic bias $H_t(a)$ maps repetition and alternation to positive and negative signs, respectively. To represent these action-specific biases that are independent of external state and reward history, the equation for the mixture policy incorporates additional terms like so:

$$\pi_t(s_t, a) = \frac{\exp\{(Q_t(s_t, a) + H_t(a) + B(a))/\tau\}}{\sum_{a^*}\exp\{(Q_t(s_t, a^*) + H_t(a^*) + B(a^*))/\tau\}}$$

Adding complexity in both learning and action bias and hysteresis

The primary model comparison here crossed factors for value-based learning (with first character "X", "0", "1", or "2" for the model label), constant bias ("C"), n-back hysteresis ("N"), and exponential hysteresis ("E") to incrementally build 72 models that were tested for each participant as an individual (**Table 2 and Table A in S1 Text**). Note that, in the original model comparison [12], the final 7-parameter model "2CE1" was built with two generalization parameters (g_A and g_S) added to an initial 5-parameter base model "0CE1" (first adding β_R, β_1, and λ_H to the standard 2-parameter base model "0" with only learning rate α and temperature τ). Unlike the original factorial model comparison, the present model comparison was more exhaustive for biases rather than reduced variants of GRL or alternative learning algorithms. Hence the bias and hysteresis factors were presently crossed with the limited cases of no learning ("X") ($\alpha = g_A = g_S = 0$), basic RL ("0") ($g_A = g_S = 0$), 1-parameter GRL ("1") ($g_A = \min\{0, g_S\}$, $-1 \leq g_S \leq 1$), and 2-parameter GRL ("2") ($-1 \leq g_A \leq 0$, $-1 \leq g_S \leq 1$).

The binary factor of constant bias was implemented as a lateral bias β_R (where a positive sign is arbitrarily rightward). Hysteresis, the next main factor, was further subdivided between exponential and n-back hysteresis as parametric and nonparametric alternatives, respectively. A model with N-back hysteresis included independent weights β_n for each of N total previous actions (the final number in the label such as the "1" in 2CN1 for 1-back), where each signed weight corresponds to a bias in favor of repetition ($\beta_n > 0$) or alternation ($\beta_n < 0$) of the respective previous action. The alternative of parametric hysteresis featured exponential decay (e.g., 2CE1) but could also include up to two additional degrees of freedom (e.g., up to 2CE3) for nonparametric weights on the most recent previous actions—that is, n-back and exponential hysteresis combined (cf. regression analyses in [17,20,58–61]).

Within each data set (i.e., the 3-T Face/House ("FH") version or the 7-T Color/Motion ("CM") version), the first step of the original analysis [12] entailed dividing participants into three subgroups according to model-independent performance on the task [18] as well as the results of model fitting [21]. A subset of participants was initially set aside as the "Good learner" ("G") group (FH: $n = 31/47$, CM: $n = 16/22$) if choice accuracy was significantly greater than the chance level of 50% for a given individual ($p < 0.05$). The remaining

participants—for whom the null hypothesis of chance accuracy could not be rejected at the individual level ($p > 0.05$)—were further subdivided between the "Poor learner" ("P") group (FH: $n = 9/47$, CM: $n = 5/22$) and the "Nonlearner" ("N") group (FH: $n = 7/47$, CM: $n = 1/22$) according to whether or not an RL or GRL model (including bias and hysteresis) could yield a significant improvement in goodness of fit relative to the pure bias-and-hysteresis model XCE1, which is nested within the full 2CE1 model adding GRL but has no sensitivity to reward or its omission.

Whereas the original model comparison [12] emphasized variants of GRL with associative or discriminative generalization and permuted these factors accordingly, the presently emphasized factors of action bias and hysteresis had been assumed a priori and fixed with three parameters for constant bias and exponential decay of the hysteresis trace. Although the original results were in favor of the 7-parameter 2CE1 model, these conclusions were drawn from only one perspective with fixed assumptions for action bias and hysteresis. That is, two new parameters for action and state generalization (g_A, g_S) were previously justified as additions to a 5-parameter base model 0CE1 starting with two parameters for basic RL (α, τ), one for constant bias (β_R), and two for exponential hysteresis (β_I, λ_H). The 3-parameter adjunct ("-CE1") was hypothesized to retain the most explanatory power post-correction in the present model comparison as well—even as various simpler and more complex alternatives were now being tested for due diligence.

Across all five participant groups from both data sets, the model comparison here established that the best-performing models featured not only GRL (for actual learners) but also constant bias and exponential hysteresis (FH-G: 2CE1, FH-P: 1CE3, FH-N: XCE2, CM-G: 2CE1, CM-P: 1CE2)—even after correcting for model complexity according to the Akaike information criterion with correction for finite sample size (AICc) [62,63] (**Figs 2A and 3A and Tables B-F in S1 Text**). Furthermore, at the individual level, 87% of participants exhibited significant effects of some kind of action-specific bias or hysteresis (FH: $n = 41/47$, CM: $n = 19/22$) (**Figs 2B and 3B and Figs Kd and Ld in S1 Text**).

With regard to correspondence between this bias-centric model comparison and the original learning-centric model comparison [12], individual Good learners were again always best fitted by a learning model (FH: $n = 31/31$, CM: $n = 16/16$), whereas Nonlearners were again always best fitted by a nonlearning model with nothing more than action bias or hysteresis (FH: $n = 0/7$, CM: $n = 0/1$). The boundary case of the Poor-learner group was mostly but not always in the direction of a learning model as opposed to a nonlearning model (FH: $n = 6/9$, CM: $n = 4/5$). Nevertheless, the original group assignments were retained here not only for consistency but also in consideration of the lack of a full factorial design with respect to GRL here (originally 11 models rather than 3).

As hypothesized for bias and hysteresis parameters, Nonlearners and even Poor learners showed greater gains in model performance than Good learners, but Good learners still benefited significantly as well. The Poor-learner and Nonlearner groups actually suggested greater explanatory power from additional hysteresis parameters (even over a third learning parameter): The best fits were from the 1CE3 and 1CE2 models for Poor learners and XCE2 for Nonlearners. Yet, in the interest of a universal model that is both parsimonious and straightforward, the 2CE1 model and the CE1 adjunct remained preferred overall for the present purposes because the Good-learner groups, which both favored 2CE1, are more reliable and more essential as evidence for a mixture of experts and nonexperts. These results and many others that follow confirmed that the original group assignments from the learning-centric model comparison remain applicable with reanalysis from this bias-centric perspective.

Although a simpler alternative nested within the 7-parameter 2CE1 model may provide a decent account for some individuals, this moderately complex model in itself provided the

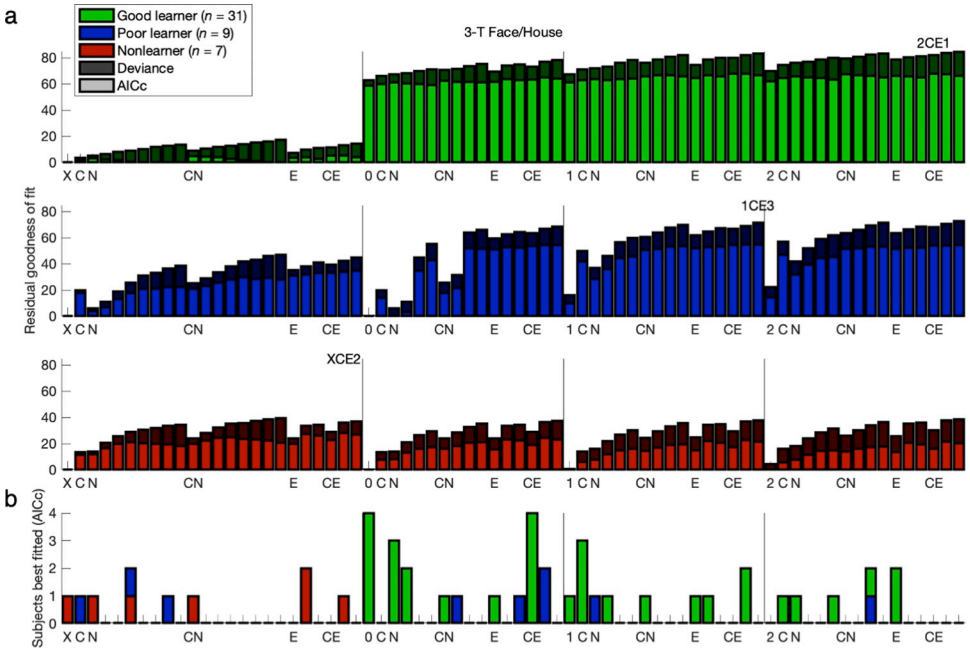

Fig 2. Model comparison: 3-T Face/House version. The ordering of the models here corresponds to the ordering in Table 2 and Table A in S1 Text. As before, the model begins with "X-", "0-", "1-", or "2-" for no learning, basic RL, 1-parameter GRL, or 2-parameter GRL. A subsequent "C" denotes constant bias, and "N" or "E" represents *n*-back or exponential hysteresis, respectively, while incrementally adding a step back to the *n*-back horizon with each successive model within a hysteresis category (e.g., the rightmost models 2CE1, 2CE2, and 2CE3). **(a)** Shown for each model is average goodness of fit relative to the null chance model ("X") with (light bars) and without (light and dark bars combined) a penalty for model complexity according to the corrected Akaike information criterion (AICc). With the addition of action bias and hysteresis parameters alongside GRL, Poor learners (blue bars) and Nonlearners (red bars) revealed the greatest gains in model performance, but Good learners (green bars) benefited significantly as well. The best-performing models (written above each plot) featured not only GRL for the actual learners but also constant bias and exponential hysteresis for all (FH-G: 2CE1, FH-P: 1CE3, FH-N: XCE2; see Fig 3 for CM-G: 2CE1, CM-P: 1CE2). For the most essential Good-learner group, the originally preferred 2CE1 model was validated as preferable to both simpler and more complex alternatives for the specification of bias and hysteresis or lack thereof. A more positive residual corresponds to a superior fit. **(b)** Counts of the participants best fitted by each model according to the AICc are plotted with separation of Good learners, Poor learners, and Nonlearners. At the individual level, 87% of participants across both data sets exhibited significant effects of some kind of action bias or hysteresis. The 7-parameter 2CE1 model—complementing 2-parameter GRL with constant bias and 2-parameter exponential hysteresis—accommodates heterogeneity in both learning and action-specific effects across individuals, leaving 64% best fit by 2CE1 or one of its nested models rather than other *n*-back or *n*-back-plus-exponential models.

https://doi.org/10.1371/journal.pcbi.1011950.g002

most parsimonious account for the greatest proportion of heterogeneous participants—and especially so among those who learned well. Conversely, the lesser overall performance of the 8- and 9-parameter models argues against an explanation reduced to mere overfitting. While omitting additional *n*-back degrees of freedom, the 2-parameter specification for exponential hysteresis was sufficiently flexible to best fit (post-correction) 64% of the heterogeneity across participants with nested models (FH: $n = 28/47$, CM: $n = 16/22$). As for the nonparametric equivalent in total degrees of freedom, substituting 2-back hysteresis (i.e., 2CN2) in lieu of the decay parameter would accommodate only 54% of this heterogeneity (FH: $n = 24/47$, CM: $n = 13/22$) in addition to providing a worse fit overall.

Having selected 2CE1 (and XCE1) with a large-scale comparison of 72 models, the most relevant subsets of eight models were rearranged for a follow-up comparison—namely, 2, 2N1, 2N2, 2E1, 2C, 2CN1, 2CN2, and 2CE1 (4 to 7 parameters) for the two learner groups and X, XN1, XN2, XE1, XC, XCN1, XCN2, and XCE1 (0 to 3 parameters) for the Nonlearner group

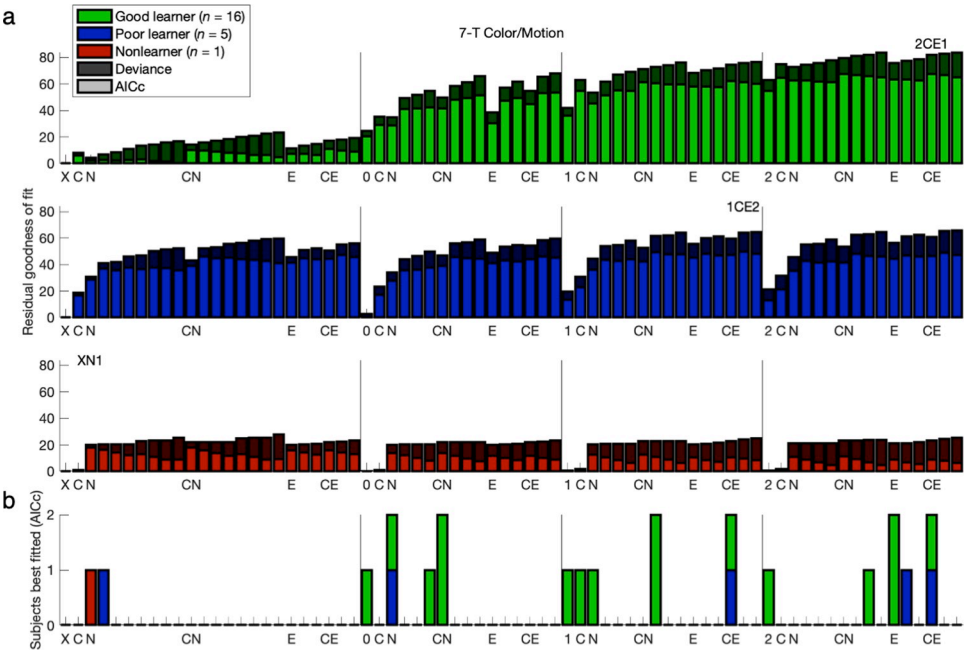

Fig 3. Model comparison: 7-T Color/Motion version. Compare to Fig 2. Results were replicated in the 7-T Color/Motion version of the experiment with a nearly identical experimental design.

https://doi.org/10.1371/journal.pcbi.1011950.g003

(**Figs 4 and 5 and Figs Ka and La in S1 Text**). Between the edge cases of the no-bias model "2" and the full model 2CE1 were another six intermediate models—that is, four nested within 2CE1 featuring exponential hysteresis (2N1, 2E1, 2C, 2CN1) and two substituting 2-back

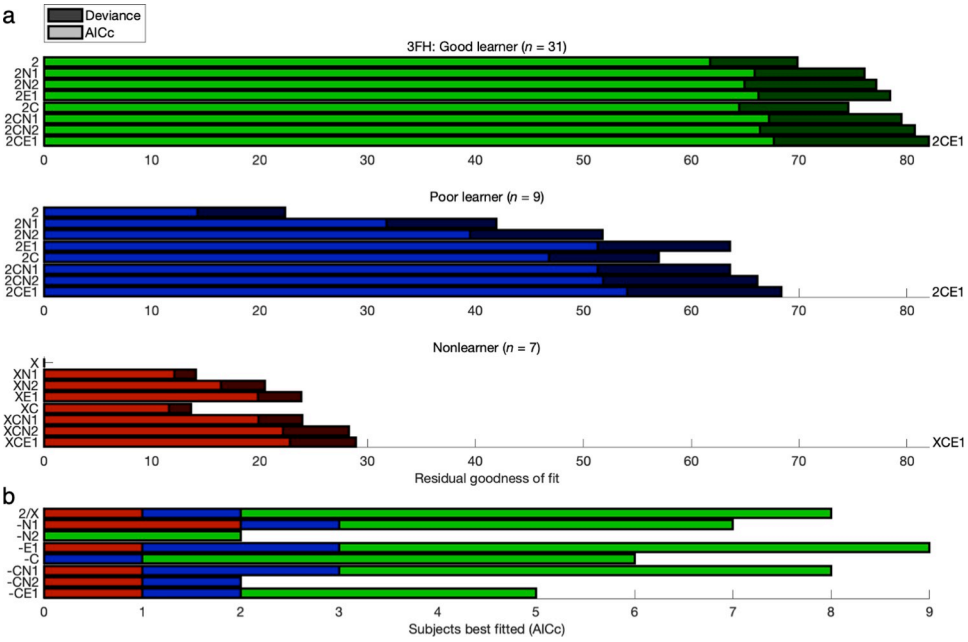

Fig 4. Reduced model comparison: 3-T Face/House version. Compare to Fig 2. The next round of comparisons focused on subsets of eight models building up to constant bias and exponential hysteresis ("-CE1"). The baseline models were 2-parameter GRL ("2") for Good and Poor learners or a random policy ("X") for Nonlearners. The evidence for best fit with the 2CE1 model is more visibly salient here (FH-G: 2CE1, FH-P: 2CE1, FH-N: XCE1; see Fig 5 for CM-G: 2CE1, CM-P: 2CN2).

https://doi.org/10.1371/journal.pcbi.1011950.g004

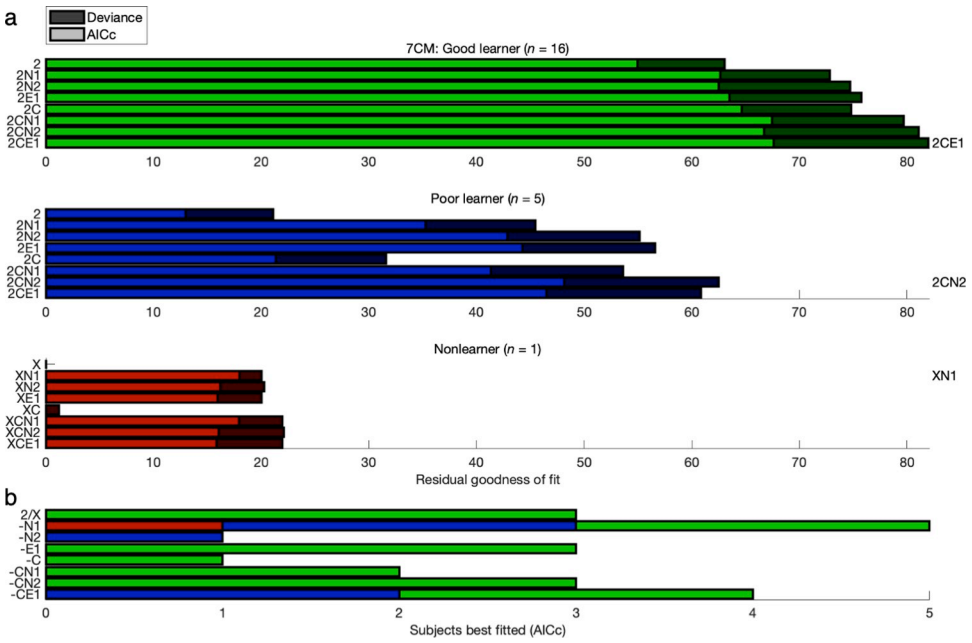

Fig 5. Reduced model comparison: 7-T Color/Motion version. Compare to Figs 3 and 4.

https://doi.org/10.1371/journal.pcbi.1011950.g005

hysteresis (2N2, 2CN2) with an equivalent number of degrees of freedom. The evidence for best fit with the 2CE1 model was more salient in this subset (FH-G: 2CE1, FH-P: 2CE1, FH-N: XCE1, CM-G: 2CE1, CM-P: 2CN2).

To again affirm the discriminability of the preferred 2CE1 model among both simpler and more complex alternatives ranging from 0 to 9 free parameters, simulated data sets were yoked to their respective empirical data sets but instead derived from individually fitted instantiations of this generative model. The simulated agent would receive input in silico according to what the respective human participant actually encountered in the session. When substituting simulated data generated by the instantiations of the 2CE1 model fitted to empirical data, the pattern of results could be replicated as expected (**Figs C, D, G, H, Kb/e, Lb/e, and M and Tables G-K in S1 Text**). Conversely, yoked simulations generated by the no-bias model "2" with only GRL—that is, a reduced model still biased toward reward maximization but unbiased with respect to action bias and hysteresis—shifted the fitting results to instead align with a learning-sans-bias model as expected (**Figs E, F, I, J, Kc/Kf, Lc/Lf, and M and Tables L-P in S1 Text**). In other words, the more complex model could be recovered from itself, and the simpler model could be recovered from itself, ruling out overfitting.

With the model comparison again (cf. [12]) pointing to the 7-parameter 2CE1 model, the individually fitted parameters of this model were verified and interpreted with reference to model-independent metrics for either action bias or learning performance (**Table 3**). The constant lateral bias β_R could be either leftward ($\beta_R < 0$) or rightward ($\beta_R > 0$), and its absolute value $|\beta_R|$ represents the weight of constant bias independent of direction—thereby resolving ambiguity between truly zero bias in the aggregate versus a distribution of substantial nonzero biases that are both positive and negative among individuals so as to cancel each other out. The initial magnitude of the exponential hysteresis bias β_1 could accommodate both repetition ($\beta_1 > 0$) and alternation ($\beta_1 < 0$), where the unsigned weight $|\beta_1|$ represents either form in the 1-back hysteretic bias. Furthermore, the model's overall level of bias—or at least 0-back and

Table 3. Parameters of the 2CE1 model. Fitted parameters for the preferred 2CE1 model are listed for each participant group based on learning performance. To characterize the dimensions of distinct behavioral profiles for each participant, the signs of individual fits are categorized as "discriminative" ($-1 \leq g_A < 0$) or "none" ($g_A = 0$) for action generalization; "discriminative" ($-1 \leq g_S < 0$), "none" ($g_S = 0$), or "associative" ($0 < g_S \leq 1$) for state generalization; "leftward" or ($\beta_R < 0$) "rightward" ($\beta_R > 0$) for constant bias; and "alternation" ($\beta_1 < 0$) or "repetition" ($\beta_1 > 0$) for hysteretic bias. Also listed are metrics for absolute constant bias $|\beta_R|$, absolute hysteretic bias $|\beta_1|$, and overall bias $|\beta_R|+|\beta_1|$, which is inversely related to the probability of a correct response ($p < 0.05$). The residual deviance D_{df} (with degrees of freedom in the subscript) corresponds to the 2CE1 model's improvement in fit relative to either the XC model with only constant bias or the complete nonlearning model XCE1 adding exponential hysteresis. Standard deviations are listed in parentheses below corresponding means.

	3-T Face/House			7-T Color/Motion		
2D GRL + Con + Exp (2CE1)	**Good learner**	**Poor learner**	**Non-learner**	**Good learner**	**Poor learner**	**Non-learner**
n	31	9	7	16	5	1
Learning rate α	0.517 (0.242)	0.269 (0.339)	0.483 (0.345)	0.555 (0.345)	0.540 (0.353)	0.372
Action generalization g_A	-0.355 (0.367)	-0.321 (0.376)	-0.787 (0.357)	-0.535 (0.393)	-0.551 (0.482)	-1.000
Discriminative : None	21 : 10	6 : 3	7 : 0	13 : 3	4 : 1	1 : 0
State generalization g_S	-0.184 (0.344)	0.367 (0.535)	0.359 (0.887)	-0.239 (0.390)	0.257 (0.819)	1.000
Disc. : None : Associative	18 : 9 : 4	1 : 2 : 6	2 : 0 : 5	11 : 1 : 4	1 : 1 : 3	0 : 0 : 1
Softmax temperature τ	0.698 (0.464)	0.737 (0.565)	3.066 (0.724)	0.700 (0.343)	1.298 (0.782)	2.157
Rightward bias β_R	0.113 (0.354)	0.160 (0.185)	0.391 (0.855)	0.167 (0.240)	0.245 (0.360)	-0.435
Leftward : Rightward	12 : 19	2 : 7	2 : 5	2 : 14	1 : 4	1 : 0
Repetition bias: Initial magnitude β_1	-0.066 (0.235)	-0.133 (0.438)	-0.169 (1.034)	-0.130 (0.153)	-0.393 (0.949)	-1.278
Alternation : Repetition	21 : 10	4 : 5	4 : 3	13 : 3	3 : 2	1 : 0
Repetition bias: Inverse decay rate λ_H	0.543 (0.371)	0.578 (0.404)	0.456 (0.421)	0.659 (0.318)	0.485 (0.403)	0.000
Constant bias $\|\beta_R\|$	0.196 (0.314)	0.191 (0.149)	0.714 (0.561)	0.207 (0.204)	0.305 (0.298)	0.435
Hysteretic bias $\|\beta_1\|$	0.171 (0.172)	0.228 (0.392)	0.868 (0.472)	0.152 (0.130)	0.717 (0.672)	1.278
Overall bias $\|\beta_R\|+\|\beta_1\|$	0.367 (0.449)	0.419 (0.396)	1.582 (0.969)	0.358 (0.276)	1.021 (0.584)	1.713
Constant (XC): Res. dev. D_6	78.56	48.67	16.75	73.97	42.15	22.28
C + Exponential (XCE1): D_4	70.55	28.94	1.46	64.82	10.31	1.51

https://doi.org/10.1371/journal.pcbi.1011950.t003

1-back bias while overlooking the decaying remainder—could be quantified as $|\beta_R|+|\beta_1|$ for a metric.

To more rigorously test for effects of action bias and hysteresis even in the presence of competing effects of value-based learning, Nonlearners are excluded from many of the analyses that follow. Confirming parameter validity across both Good and Poor learners, the rightward bias β_R was correlated with the probability of performing the right-hand action (FH: $r = 0.556$, $t_{38} = 4.13$, $p < 10^{-4}$; CM: $r = 0.640$, $t_{19} = 3.63$, $p < 10^{-3}$). Likewise, the repetition bias β_1 was correlated with the probability of repeating the previous action regardless of state (FH: $r = 0.769$, $t_{38} = 7.40$, $p < 10^{-8}$; CM: $r = 0.660$, $t_{19} = 3.83$, $p < 10^{-3}$). Given the exclusively right-handed participants in this study, the majority were expected to exhibit a net rightward bias ($\beta_R > 0$) like that even captured within the subgroups based on learning performance (FH-G: $M = 0.113$, $t_{30} = 1.78$, $p = 0.043$; FH-P: $M = 0.160$, $t_8 = 2.60$, $p = 0.016$; FH-N: $M = 0.391$, $t_6 = 1.21$, $p = 0.136$; CM-G: $M = 0.167$, $t_{15} = 2.78$, $p = 0.007$; CM-P: $M = 0.245$, $t_4 = 1.52$, $p = 0.102$).

If action bias and hysteresis were omitted as is typically the case, estimation bias and other distortions of learning parameters would arise when forced to simultaneously fit these parallel phenomena that are otherwise unaccounted for. The necessity of the extra parameters could also be validated in silico with parameter recovery or lack thereof when simulating with or without bias parameters, respectively (**Fig N in S1 Text**). As compared with successfully recovering parameters of the full bias-and-hysteresis model 2CE1 (FH-G: α: $r = 0.759$, $p < 10^{-6}$; g_A: $r = 0.731$, $p = 10^{-6}$; g_S: $r = 0.725$, $p = < 10^{-5}$; τ: $r = 0.668$, $p < 10^{-4}$; β_R: $r = 0.624$, $p = 10^{-4}$; β_1: $r = 0.876$, $p < 10^{-10}$; λ_H: $r = 0.463$, $p = 0.004$; FH-P: α: $r = 0.841$, $p = 0.002$; g_A: $r = 0.853$, $p = 0.002$; g_S: $r = 0.819$, $p = 0.003$; τ: $r = 0.306$, $p = 0.212$; β_R: $r = 0.824$, $p = 0.003$; β_1: $r = 0.725$, $p = 0.014$; λ_H: $r = 0.666$, $p = 0.025$; CM-G: α: $r = 0.638$, $p = 0.004$; g_A: $r = 0.472$, $p = 0.033$; g_S: $r = 0.621$, $p = 0.005$; τ: $r = 0.697$, $p = 10^{-3}$; β_R: $r = 0.717$, $p < 10^{-3}$; β_1: $r = 0.588$, $p = 0.008$; λ_H: $r = 0.448$, $p = 0.041$; CM-P: α: $r = 0.786$, $p = 0.058$; g_A: $r = 0.866$, $p = 0.029$; g_S: $r = 0.891$, $p = 0.021$; τ: $r = 0.885$, $p = 0.023$; β_R: $r = 0.974$, $p = 0.003$; β_1: $r = 0.856$, $p = 0.032$; λ_H: $r = 0.996$, $p < 10^{-3}$), recovery of the learning parameters from 2CE1 with the no-bias model "2" was generally less robust for all learners and especially insufficient—even failing to recover—for the Poor-learner group more characterized by action biases that outweigh and obscure confounded learning processes (FH-G: α: $r = 0.291$, $p = 0.056$; g_A: $r = 0.535$, $p = 10^{-3}$; g_S: $r = 0.744$, $p < 10^{-6}$; τ: $r = 0.658$, $p < 10^{-4}$; FH-P: α: $r = 0.430$, $p = 0.124$; g_A: $r = 0.172$, $p = 0.329$; g_S: $r = 0.418$, $p = 0.131$; τ: $r = 0.374$, $p = 0.161$; CM-G: α: $r = 0.683$, $p = 0.002$; g_A: $r = 0.592$, $p = 0.008$; g_S: $r = 0.604$, $p = 0.007$; τ: $r = 0.690$, $p = 0.002$; CM-P: α: $r = 0.716$, $p = 0.087$; g_A: $r = 0.631$, $p = 0.127$; g_S: $r = 0.995$, $p < 10^{-3}$; τ: $r = 0.796$, $p = 0.054$).

The deficiencies of a model limited to only learning are especially noteworthy in this contrived environment with experimental controls regulating the reward schedule such that spurious confounds between effects of learning and effects of bias and hysteresis have been mitigated by design. The proof of concept in this extreme case with unnatural controls suggests an even more pressing need for this framework for applications in less controlled laboratory settings as well as natural settings in the real world. Elsewhere without such experimental control via deliberate counterbalancing that would otherwise impose symmetric structure in the environment as well as individual trajectories within it, there would be even greater susceptibility to parameter distortion if bias parameters were omitted.

Action bias and hysteresis versus learning performance

In keeping with the previous point about idiosyncratic environments, the statistics of a given task environment must be considered to set reference points for quantifying and interpreting truly action-specific components of variance. While triple dissociation of bias, hysteresis, and learning is generally nontrivial for a short sequence of active states, this challenge can be exacerbated even more so by class imbalance depending on the temporal statistics of states, actions, and rewards. In arriving at a fully interpretable quantitative model amenable to individual differences, the challenge was first met here by a hierarchically counterbalanced experimental design that was tightly controlled within and across sessions.

Regarding the constant lateral bias, available rewards were thus evenly distributed between left-hand and right-hand actions all throughout the experiment. Hence an omniscient optimal agent with perfect 100% accuracy would be guaranteed to produce an even 50% probability of a left- or right-hand action. This was not the case for hysteresis, however.

In contrast, that same agent would produce an uneven 66.7% probability of action alternation as a byproduct of choosing the optimal actions here. This incidental asymmetry can superficially mimic an internal alternation bias while a learner actually responds to the external structured sequence of four randomly rotating states. (States were never repeated in

consecutive trials, and of the three remaining states, only one from the complementary category would reward the action just performed in a given state for the block—resulting in two-thirds or 66.7% alternation.) Note that a naïve policy with a 100% probability of alternation irrespective of state would nonetheless produce chance accuracy at 50% by design. Such ambiguity for a raw, model-independent measure again underscores the need for comprehensive computational modeling that accounts for multiple implicit effects simultaneously.

To the extent that the forces of bias and learning compete with each other to drive behavior, an inverse relation was expected between learning performance and the weight of action bias and hysteresis. Again omitting Nonlearners, overall bias $|\beta_R|+|\beta_I|$ in actual learners was inversely correlated with accuracy as the probability of choosing the correct action (FH: r = -0.290, t_{38} = 1.87, p = 0.035, r_S = -0.374, p = 0.009 for monotonicity; CM: r = -0.472, t_{19} = 2.33, p = 0.015, r_S = -0.605, p = 0.002 for monotonicity). This inverse relation between modeled bias and objective performance was monotonic across not only all learners but also the alternation-bias group specifically (FH: r = -0.383, t_{23} = 1.99, p = 0.029, r_S = -0.475, p = 0.009 for monotonicity; CM: r = -0.453, t_{14} = 1.90, p = 0.039, r_S = -0.618, p = 0.006 for monotonicity), demonstrating that bias as extracted with modeling was not confounded with alternation that may incidentally result from pursuing reward. (See next section for more detail about the alternation-bias group.)

To complement the initial quantitative model comparison for overall goodness of fit, a series of posterior predictive checks followed for evidence of bias and hysteresis with qualitative falsification of the null hypotheses in nested models [26–28]. The same technique had been used previously to falsify basic RL against GRL [12]. Each check entailed juxtaposition of empirical behavior and the behavior simulated by GRL models that, while holding a fixed assumption of two new learning parameters for generalization, are incrementally tested with up to three more action-bias parameters.

First separating groups on the basis of learning performance, a binary model comparison could illustrate some fundamental limitations of the pure GRL model "2" with no bias as opposed to the final 2CE1 model with three parameters for constant bias and exponential hysteresis. (The intermediate models between these 4- and 7-parameter end points are investigated in greater depth later.) Posterior predictive checks for these two models were tested against empirical results for not only the probability of a correct (versus incorrect) action—as is standard for a learning paradigm—but also the probability of a right-hand (versus left-hand) action and the probability of a repeated (versus alternated) action independent of state.

From a naïve perspective it would appear that, by qualitatively capturing the probability of a correct choice across levels of learning performance (FH-G: M = 12.8%, t_{30} = 13.13, $p < 10^{-13}$; FH-P: M = 0.1%, $p > 0.05$; FH-N: M = 0.1%, $p > 0.05$; CM-G: M = 12.3%, t_{15} = 8.75, $p = 10^{-7}$; CM-P: M = -0.2%, $p > 0.05$) in silico as well (FH-G: $p < 0.05$; FH-P: $p > 0.05$; FH-N: $p > 0.05$; CM-G: $p < 0.05$; CM-P: $p > 0.05$) (**Figs 6A/6D and 7A/7D and Fig Oa/d in S1 Text**), the 4-parameter GRL model "2" with no bias seemingly accounts for human behavior comparably to the 7-parameter 2CE1 model expanded with action bias and hysteresis. However, the shortcomings of a purely learning-based account can be revealed even in 0-back and 1-back action-specific effects. Remarkably, these action-specific effects (**Figs 6E–6F and 7E–7F**) are quite substantial in effect size as compared with the value-based effects (**Figs 6D and 7D**) typically and most intuitively emphasized in a paradigm for active learning.

Across these right-handed participants, all five groups in the aggregate performed the right-hand action more often (FH-G: M = 1.8%, t_{30} = 2.11, p = 0.022; FH-P: M = 9.3%, t_8 = 3.99, p = 0.002; FH-N: M = 5.1, t_6 = 1.54, p = 0.088; CM-G: M = 4.8%, t_{15} = 3.21, p = 0.003; CM-P: M = 9.9%, t_4 = 2.36, p = 0.039) (**Figs 6B/6E and 7B/7E and Fig Ob/Oe in S1 Text**), and greater or marginally greater rightward bias was observed in Poor learners and Nonlearners relative to

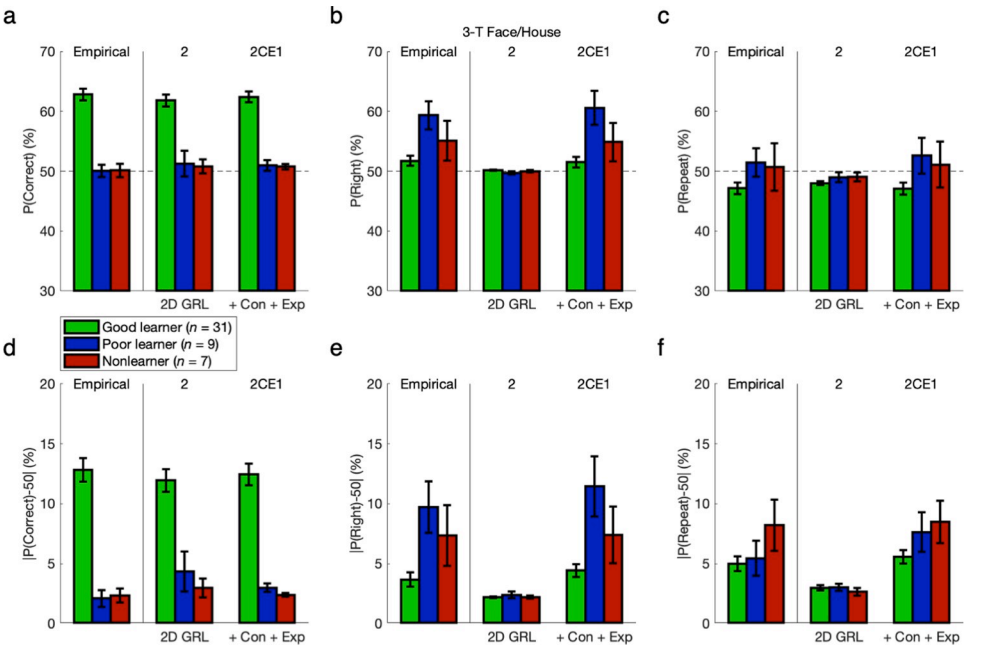

Fig 6. Action bias and hysteresis versus learning performance: 3-T Face/House version. To compare the pure GRL model ("2") with the final 2CE1 model adding three parameters for constant bias and exponential hysteresis, simulated data sets from each model were yoked to their respective empirical data sets. Posterior predictive checks were tested for the probability of a correct action, the probability of a right-hand action, or the probability of a repeated action independent of state. **(a)** If only examining accuracy in terms of correct choices for maximizing reward, the shortcomings of the reduced model without bias are not so obviously apparent at first. **(b)** Upon considering action bias, these right-handed individuals mostly had a tendency to select the right-hand action ($p < 0.05$). Whereas the 2CE1 model could account for this effect with a constant lateral bias ($p < 0.05$), the reduced model could not ($p > 0.05$). **(c)** Regarding the probability of repetition versus alternation, note that 100% accuracy would produce 66.7% alternation for the present experimental design, but 100% alternation would still produce 50% accuracy. The Good-learner group exhibited a tendency to alternate in the aggregate as expected ($p < 0.05$), whereas the Poor-learner and Nonlearner groups did not ($p > 0.05$). Only the 2CE1 model featuring exponential hysteresis could match this pattern with quantitative precision. **(d-f)** Independent of direction, absolute differences from the chance level of 50% reveal the full extent of the action-specific components of variance, which are as substantial as the effects of reward typically emphasized in active learning. For fitting the probability of a right-hand action or a repeated action, a margin of roughly 2% for pure GRL was insubstantial in comparison. Error bars indicate standard errors of the means.

https://doi.org/10.1371/journal.pcbi.1011950.g006

Good learners (FH-PG: $M = 7.6\%$, $t_{38} = 3.80$, $p < 10^{-3}$; FH-NG: $M = 3.3\%$, $t_{36} = 1.43$, $p = 0.081$; CM-PG: $M = 5.2\%$, $t_{19} = 1.47$, $p = 0.079$). Hence this measure of absolute lateral bias $|P(Right)\text{-}50\%|$ was also greater in Poor learners and Nonlearners (FH-PG: $M = 6.0\%$, $t_{38} = 3.81$, $p < 10^{-3}$; FH-NG: $M = 3.7\%$, $t_{36} = 2.14$, $p = 0.020$; CM-PG: $M = 4.8\%$, $t_{19} = 1.51$, $p = 0.074$), which likewise held true when correlating across the continuous measure of accuracy rather than discrete participant groups (FH: $r = -0.544$, $t_{38} = 4.00$, $p = 10^{-4}$; CM: $r = -0.540$, $t_{19} = 2.80$, $p = 0.006$). Whereas the full 2CE1 model could replicate all of these effects ($p < 0.05$), the reduced GRL model could not ($p > 0.05$). As a reflection of individual-specific class imbalance or overfitting in the absence of constant bias, a roughly 2% margin was apparent in the absolute difference between the reduced model's right-hand probability and the chance level of 50% (**Figs 6E and 7E**). Yet this margin was insubstantial in comparison to the true effect sizes of constant bias that were quantitatively matched by only the full model.

Note again that 100% accuracy in this contrived environment would produce 66.7% alternation because of rotating states, but 100% alternation would produce 50% accuracy. The interpretation of this raw measure is thus confounded between effects of reward and hysteresis, but in keeping with the statistics of the environment, the Good-learner groups did exhibit a tendency to

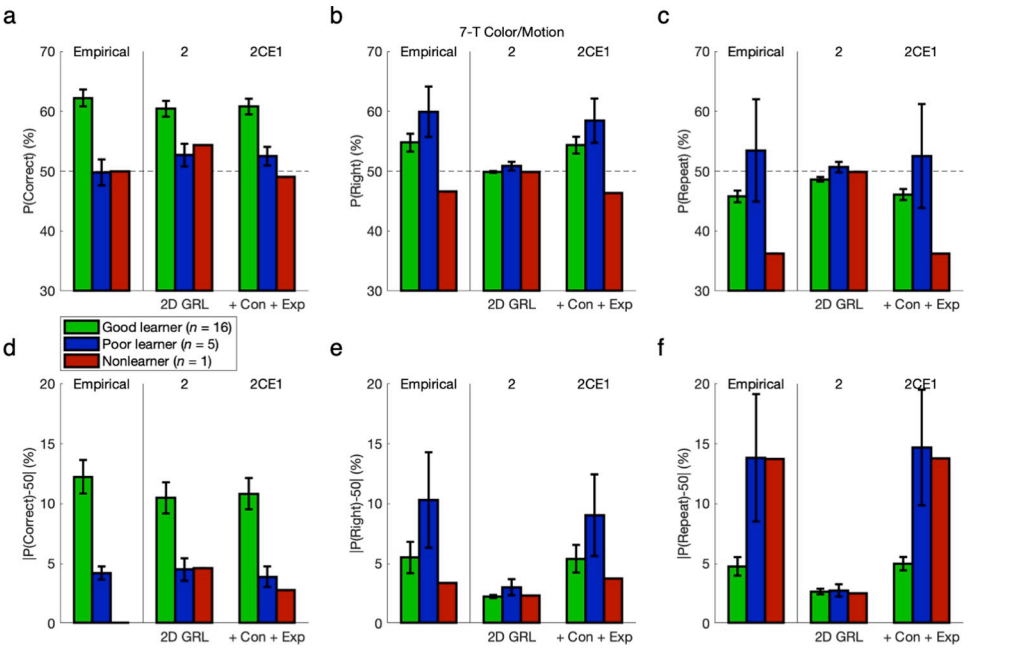

Fig 7. Action bias and hysteresis versus learning performance: 7-T Color/Motion version. Compare to Fig 6. Results were replicated in the 7-T Color/Motion version of the experiment.

https://doi.org/10.1371/journal.pcbi.1011950.g007

alternate in the aggregate while the Poor-learner and Nonlearner groups did not (FH-G: M = -2.9%, $t_{30} = 2.94$, $p = 0.003$; FH-P: $M = 1.5\%$, $p > 0.05$; FH-N: $M = 0.8\%$, $p > 0.05$; CM-G: M = -4.2%, $t_{15} = 4.34$, $p < 10^{-3}$; CM-P: $M = 3.5\%$, $p > 0.05$) (**Figs 6C/6F and 7C/7F and Fig Oc/f in S1 Text**). In contrast, the absolute repetition-or-alternation frequency $|P(Repeat)\text{-}50\%|$ was significantly greater than chance for all subgroups (FH-G: $M = 5.0\%$, $t_{30} = 8.11$, $p < 10^{-8}$; FH-P: $M = 5.5\%$, $t_8 = 3.73$, $p = 0.003$; FH-N: $M = 8.2\%$, $t_6 = 3.84$, $p = 0.004$; CM-G: $M = 4.8\%$, $t_{15} = 6.15$, $p < 10^{-5}$; CM-P: $M = 13.8\%$, $t_4 = 2.60$, $p = 0.030$). Relative to Good learners, Nonlearners exhibited even greater deviation from chance with repetition or alternation ($M = 3.2\%$, $t_{36} = 1.97$, $p = 0.028$), as did the Poor learners of at least the second data set ($M = 9.1\%$, $t_{19} = 2.89$, $p = 0.005$). The latter trend held true for the second data set with marginal significance for the continuous measure of accuracy as well (r = -0.312, $t_{19} = 1.43$, $p = 0.084$). Only the 7-parameter model could match net 1-back effects with quantitative precision (FH-G: $p < 0.05$; FH-P: $p > 0.05$; FH-N: $p > 0.05$; CM-G: $p < 0.05$; CM-P: $p > 0.05$), and qualitative falsification of the pure GRL model for such hysteretic effects was to be found in follow-up analyses disambiguating effects of reward and hysteresis. Owing to this disambiguation, the model-based results that follow are more reliable than these model-independent measures for inference about actual hysteresis per se.

Different forms of action bias and hysteresis

The 2CE1 model should accommodate the idiosyncrasies of individual participants with respect to not only GRL, which has already been demonstrated [12], but also action bias and hysteresis. Based on parameter fits, Good and Poor learners were combined and then reclassified according to the directionality of either constant bias or hysteretic bias—that is, leftward ($\beta_R < 0$) versus rightward ($\beta_R > 0$) or alternation ($\beta_1 < 0$) versus repetition ($\beta_1 > 0$). Nonlearners were again omitted for more rigorous testing of biases in the presence of actual learning. Each posterior predictive check was extended to the eight models previously highlighted in the

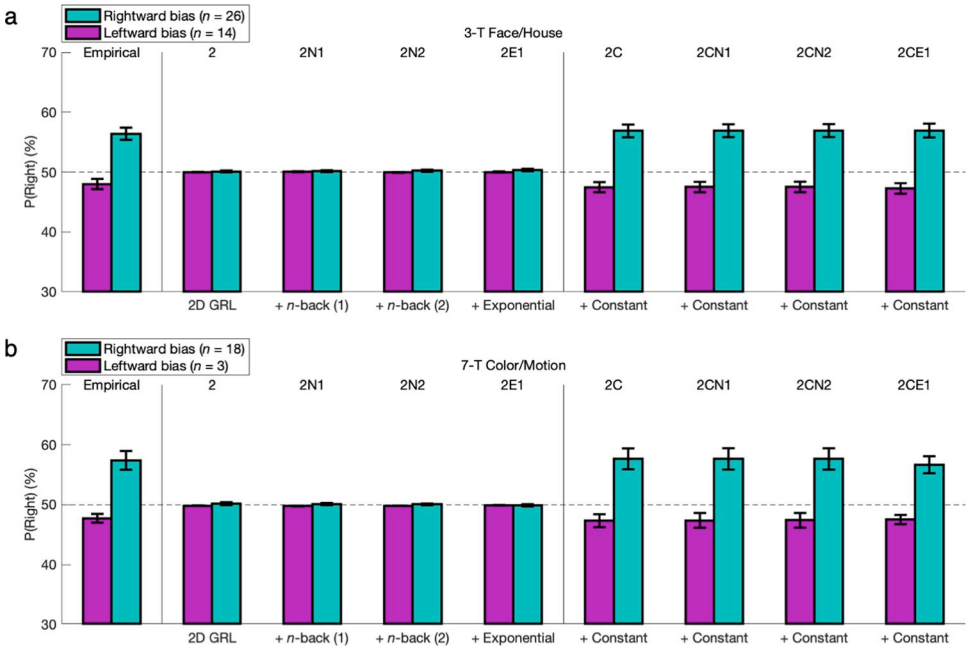

Fig 8. Constant bias. (a) Based on individual fits of the 2CE1 model, Good and Poor learners were combined and then reclassified according to whether the constant lateral bias was a leftward bias ($\beta_R < 0$) (magenta bars) or a rightward bias ($\beta_R > 0$) (cyan bars). The model comparison extended this posterior predictive check and others to another six intermediate models—four models nested within the 2CE1 model featuring exponential hysteresis (2N1, 2E1, 2C, 2CN1) and two models substituting 2-back hysteresis (2N2, 2CN2) but matched for degrees of freedom. For the probabilities of left or right actions, some of these right-handed people actually exhibited a contrary leftward bias; those who did exhibited a smaller absolute magnitude of bias than that of the rightward-bias group ($p < 0.05$). The models with a parameter for constant bias (2C through 2CE1) could replicate these effects ($p < 0.05$), falsifying the models that could not at all for lack of this parameter ($p > 0.05$). **(b)** Results were replicated in the 7-T Color/Motion version of the experiment.

https://doi.org/10.1371/journal.pcbi.1011950.g008

reduced model comparison—that is, incrementally building up from the no-bias model "2" with only GRL (4 parameters) to the full 2CE1 model (7 parameters). Necessity could thus be verified for every single parameter of the 2CE1 model.

Among these right-handed learners, 28% exhibited a contrary leftward bias (FH: $n = 14/40$; CM: $n = 3/21$). Those with leftward bias (FH: $M = -2.0\%$, $t_{13} = 2.29$, $p = 0.020$; CM: $M = -2.3\%$, $t_2 = 3.12$, $p = 0.045$) exhibited a smaller (or marginally smaller) absolute magnitude of bias (FH: $M = 4.2\%$, $t_{38} = 2.84$, $p = 0.004$; CM: $M = 5.1\%$, $t_{19} = 1.31$, $p = 0.103$) relative to the rightward-bias group (FH: $M = 6.4\%$, $t_{25} = 6.30$, $p < 10^{-6}$; CM: $M = 7.4\%$, $t_{17} = 4.73$, $p < 10^{-4}$) (Fig 8), but the existence of so many leftward biases among right-handed individuals is noteworthy. The models with a parameter for constant bias (2C through 2CE1) could replicate these effects ($p < 0.05$), whereas those without the parameter could not at all ($p > 0.05$). These findings falsify the naïve hypothesis that handedness might determine the direction of constant bias invariably. The unpredictable distribution of an effect as simple as laterality stands among the evidence that, in general, individual differences must be modeled without a-priori distributional assumptions—whether about a random sample of individuals or about the population from which they are drawn (see Discussion).

Bear in mind that optimal behavior results in more frequent alternation of actions in this particular setting. Conversely, naïve alternation does not result in above-chance performance for the aforementioned reasons. Despite the latter fact, behavior was hypothesized to be predisposed to alternation that is independent of states and outcomes after an agent has been

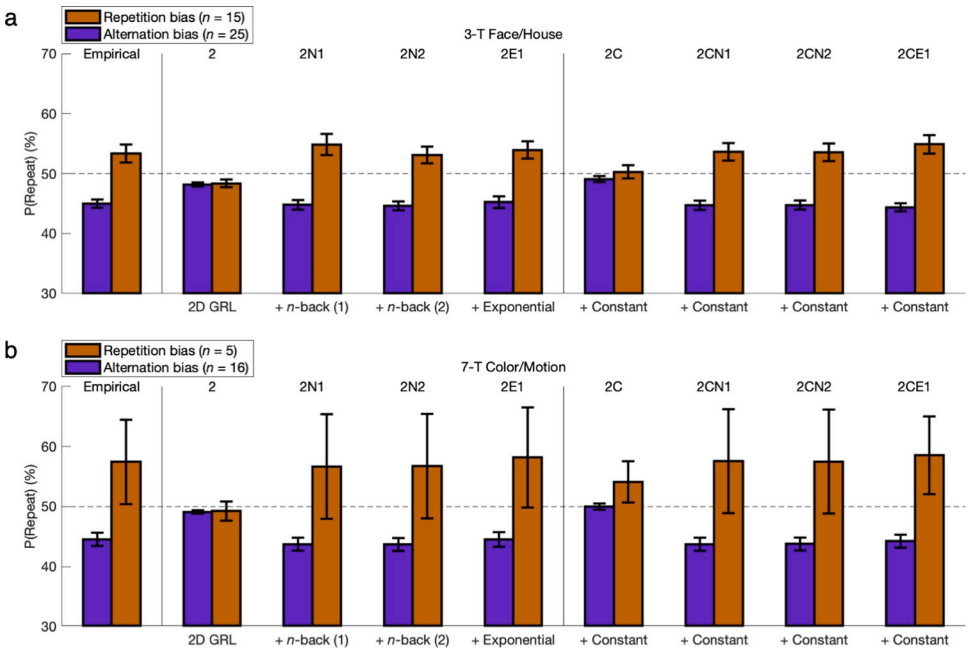

Fig 9. Hysteresis represented by the previous trial. The learners were next reclassified according to whether the hysteretic bias was an alternation bias ($\beta_1 < 0$) (violet bars) or a repetition bias ($\beta_1 > 0$) (orange bars). With some adhering to a more typical profile of first-order perseveration, the repetition-bias group did retain a substantial effect on the probability of repeating an action independent of state ($p < 0.05$). However, in keeping with second-order perseveration, the alternation-bias group actually outnumbered and outweighed in effect size the repetition-bias group ($p < 0.05$). That is, extra alternation could follow from the design feature whereby optimal behavior would more frequently result in alternating actions. In contrast to optimal alternation when appropriate for a given state, this perseverative alternation was action-specific so as to not actually improve reward-maximizing accuracy for the alternation-bias group ($p > 0.05$). The models with at least one parameter for hysteretic bias could replicate these 1-back effects ($p < 0.05$). Although the 2C model with constant bias could partially mimic action repetition with a nonsignificant trend, the models without any hysteresis parameters (2 and 2C) could not properly match the empirical 1-back effect ($p > 0.05$).

https://doi.org/10.1371/journal.pcbi.1011950.g009

alternating actions at the appropriate times due to learning that is dependent on states and outcomes. This hypothesis might initially appear at odds with the typical narrative in the RL literature emphasizing perseveration as naïve action repetition, but here, that would only represent first-order perseveration at the level of actions. At the level of policies, second-order perseveration suggests that a learner in such an environment perseverates from an expert reward-seeking policy of optimal alternation when appropriate to a nonexpert default policy of perseverative alternation whenever.

In keeping with this hypothesis, the alternation-bias group (FH: $n = 25/40$; CM: $n = 16/21$) was expected to outnumber the repetition-bias group (FH: $n = 15/40$; CM: $n = 5/21$) as well as exhibit an effect on the raw probability of alternation (FH: $M = -5.0\%$, $t_{24} = 7.32$, $p < 10^{-7}$; CM: $M = -5.4\%$, $t_{15} = 4.93$, $p < 10^{-4}$) (Fig 9). Yet reward-maximizing accuracy was not significantly higher for the alternation-bias group than for the repetition-bias group (FH: $M = 3.2\%$, $p > 0.05$; CM: $M = 2.2\%$, $p > 0.05$), confirming the action-specific nature of this bias as a nonexpert heuristic. The arrow of causality for the hypothesis of second-order perseveration primarily points from optimal alternation to perseverative alternation rather than vice versa. These results lend themselves to an analogy with the previously described cohort that was left-biased despite being right-handed, whereas there was still also a sizable repetition-bias group in which some learners instead adhered to a more intrinsic first-order perseveration effect like what has typically been reported in the literature. That is, this learning cohort could sometimes

alternate to exploit actions with high estimated reward when appropriate but still perseverated so as to repeat actions according to a more robust default repetition bias (FH: $M = 3.3\%$, $t_{14} = 2.24$, $p = 0.021$; CM: $M = 7.4\%$, $t_4 = 1.06$, $p = 0.175$; nonsignificant, but versus alternation-bias group: $M = 12.9\%$, $t_{19} = 3.06$, $p = 0.003$). Whereas the models with at least one parameter for hysteretic bias (including the simplest 2N1 model) could replicate these 1-back effects ($p < 0.05$), the models with no such parameter could not ($p > 0.05$).

Notably, the 2C model with constant bias but no hysteresis could partially mimic the repetition effect observed in the repetition-bias group (with a trending but nonsignificant result, $p > 0.05$). That is, a true action-repetition effect could be overfitted to some extent by instead representing only imbalanced base rates for actions. Although this reduced constant-only model fails to match the empirical repetition result quantitatively, there is cause for alarm in the qualitative trend that spuriously arises in both data sets. As discussed previously, the present environment represents a distinct active-learning paradigm in which such class imbalance is actually minimized—unlike most other environments with greater confounding in distributions for classes such as those of the actions per se or repetitions versus alternations. In general, omission of repetition bias may inflate estimates of constant bias with limited data if there is insufficient opportunity for repetition to be demonstrated across multiple actions. Likewise, omission of constant bias may inflate estimates of a confounded repetition effect. Conversely, omission of alternation bias may deflate estimates of constant bias because this alternation counteracts the incidental repetition of an action with a greater base rate. The different forms of bias and hysteresis all need to be accounted for comprehensively.

Psychometric modeling of the mixture policy

More quantitatively precise modeling of psychometric functions followed to examine the interface of value-based learning, action-specific effects, and the softmax function determining the mixture policy for action selection. The breadth of this mixture of experts and nonexperts integrated modular elements of basic RL, generalized RL, constant bias, hysteretic bias, and stochasticity from exploration as well as noise. As expected across all subgroups of learners, the probability of an action increased with the difference between the state-dependent action values $Q_t(s_t,a)$ learned by the GRL component of the 2CE1 model as fitted to empirical behavior (FH-L: $\beta = 1.544$, $t_{13} = 6.38$, $p = 10^{-5}$; FH-R: $\beta = 2.084$, $t_{25} = 6.74$, $p < 10^{-6}$; FH-A: $\beta = 1.682$, $t_{24} = 9.60$, $p < 10^{-9}$; FH-P: $\beta = 2.316$, $t_{14} = 4.61$, $p < 10^{-3}$; CM-L: $\beta = 0.938$, $t_2 = 2.67$, $p = 0.058$; CM-R: $\beta = 1.494$, $t_{17} = 7.20$, $p < 10^{-6}$; CM-A: $\beta = 1.443$, $t_{15} = 7.20$, $p < 10^{-5}$; CM-P: $\beta = 1.76$, $t_4 = 2.97$, $p = 0.021$) (**Figs 10 and 11**).

In determining the probability of left-hand versus right-hand actions, constant bias was derived from the logistic model in the appropriate directions for both the leftward-bias (FH: $\beta = -0.113$, $t_{13} = 2.93$, $p = 0.006$; CM: $\beta = -0.103$, $t_2 = 2.97$, $p = 0.049$) and rightward-bias (FH: $\beta = 0.265$, $t_{25} = 6.98$, $p = 10^{-7}$; CM: $\beta = 0.302$, $t_{17} = 5.08$, $p < 10^{-4}$) groups (Fig 10). The models featuring constant bias could replicate these effects with comparable psychometric functions ($p < 0.05$), whereas models without the parameter could not ($p > 0.05$).

For instead the probability of repeated versus alternated actions independent of state, hysteretic bias was derived from the logistic model in the appropriate directions for both the alternation-bias (FH: $\beta = -0.178$, $t_{24} = 5.21$, $p = 10^{-5}$; CM: $\beta = -0.220$, $t_{15} = 5.31$, $p < 10^{-4}$) and repetition-bias (FH: $\beta = 0.218$, $t_{14} = 4.79$, $p = 10^{-4}$; CM: $\beta = 0.462$, $t_4 = 1.35$, $p = 0.124$; nonsignificant, but versus alternation-bias group: $M = 0.682$, $t_{19} = 3.51$, $p = 0.001$) groups (Fig 11). The models featuring at least one parameter for hysteretic bias could replicate these 1-back effects with comparable psychometric functions ($p < 0.05$), and while models without the

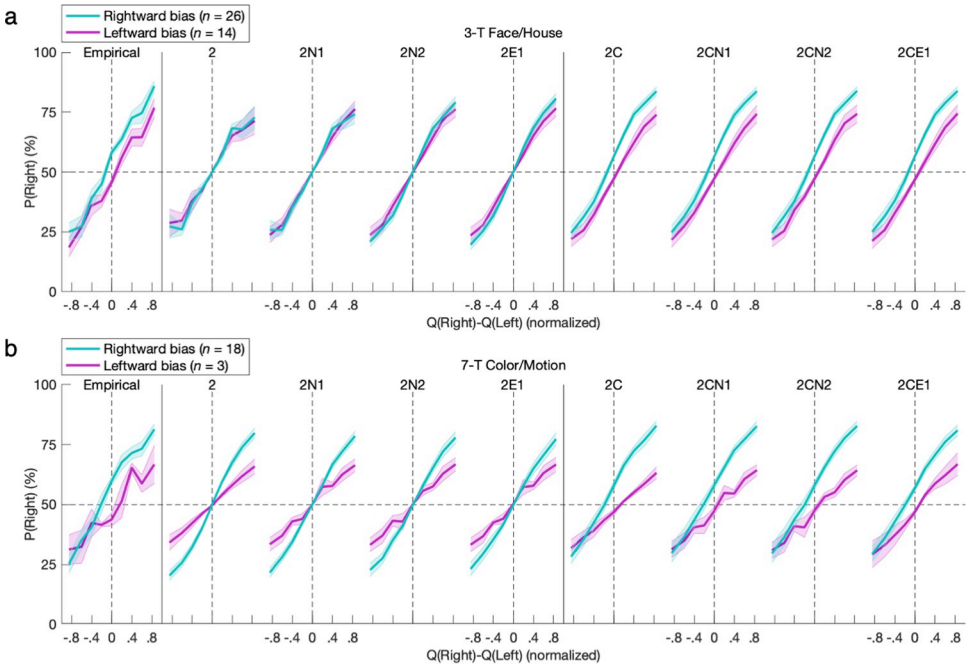

Fig 10. Psychometric modeling of constant bias. The probability of an action increased with the difference between action values $Q_t(s_t,a)$ derived from the GRL component of the 2CE1 model as fitted to empirical behavior ($p < 0.05$). Constant bias was derived from a logistic model in the appropriate directions for both the leftward-bias and rightward-bias groups ($p < 0.05$). The models featuring constant bias could replicate these effects with quantitative precision as well ($p < 0.05$), whereas models without the parameter could not ($p > 0.05$). The nine plots per row each have an identical x-axis despite omission of tick labels from every other plot for readability. Error bars indicate standard errors of the means.

https://doi.org/10.1371/journal.pcbi.1011950.g010

parameter could not ($p > 0.05$), the solitary constant bias of the 2C model does deceptively mimic repetition with a nonsignificant trend.

Dynamics of action hysteresis

The hysteresis trace of the 2CE1 model extends its temporal horizon beyond the 1-back effects examined thus far. For the preceding posterior predictive checks, the extra parameter for exponential decay could not explicitly show the full extent of its impact—showing instead only subtle quantitative improvement. If this costly free parameter were to be justified, its improvement for the model would need to also be qualitative and substantial. Considering that the 2CE1 model has already been shown to outperform both simpler and more complex implementations of hysteresis overall, the assumption of two parameters for exponential hysteresis must provide a superior parsimonious fit for effects of action history ranging from 2-back onward with an indefinite horizon. Moreover, 2-parameter exponential hysteresis outperformed n-back models for not only $n = 1$ but also $n = 2$ (2CN1 and 2CN2), establishing that it must not be only the 2-back effects but rather also 3-back and beyond that have significant weight beyond 1-back. Accordingly, hysteretic effects were explored directly up to eight trials back.

The probability of a repeated action was now conditioned on each respective action from the eight most recent trials (Fig 12; see **Fig P in** S1 Text for distributions of runs of consecutive repeats). As expected for the repetition-bias group, this probability of repeating a previous action (FH: $M = 3.3\%$, $t_{14} = 2.24$, $p = 0.021$; CM: $M = 7.4\%$, $t_4 = 1.06$, $p = 0.175$; nonsignificant,

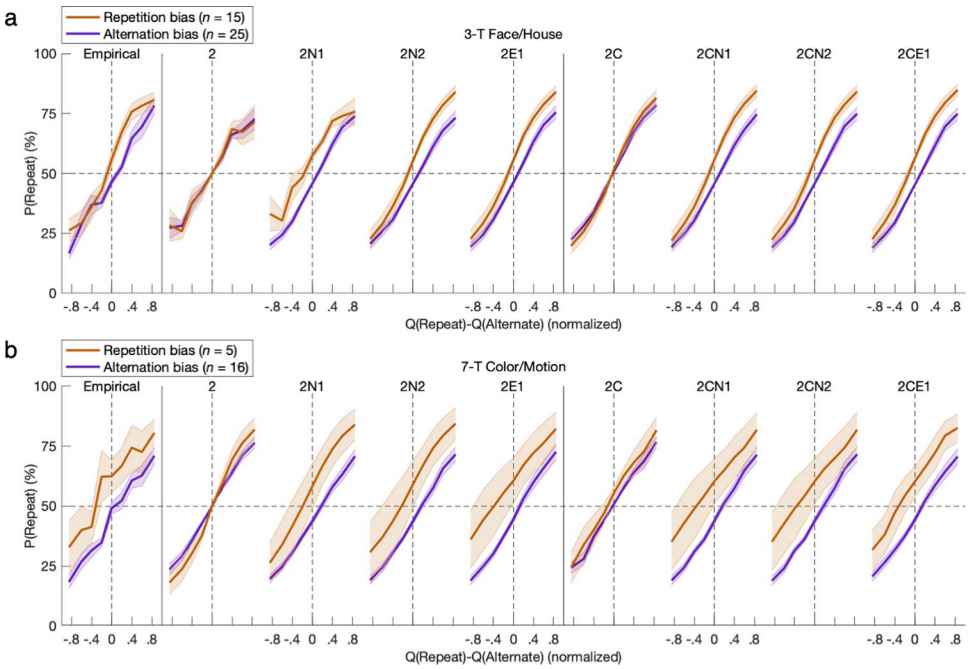

Fig 11. Psychometric modeling of hysteresis represented by the previous trial. For instead the probabilities of alternated or repeated actions, hysteretic bias was likewise derived from a GRL-based logistic model in the appropriate directions for both the alternation-bias and repetition-bias groups ($p < 0.05$). The models featuring at least one parameter for hysteretic bias could replicate these 1-back effects with comparable psychometric functions ($p < 0.05$), and while models without the parameter could not ($p > 0.05$), the 2C model could again deceptively mimic repetition with a nonsignificant trend.

https://doi.org/10.1371/journal.pcbi.1011950.g011

but versus alternation-bias group: $M = 12.9\%$, $t_{19} = 3.06$, $p = 0.003$) was elevated above chance prior to 1-back as well (FH: $M = 4.1\%$, $t_{14} = 3.39$, $p = 0.002$; CM: $M = 8.3\%$, $t_4 = 1.83$, $p = 0.070$ with marginal significance) and remained elevated. Conversely, for the alternation-bias group, this probability returned from a 1-back alternation effect (FH: $M = -5.0\%$, $t_{24} = 7.32$, $p < 10^{-7}$; CM: $M = -5.4\%$, $t_{15} = 4.93$, $p < 10^{-4}$) to the chance level prior to 1-back (FH: $M = -0.3\%$, $p > 0.05$; CM: $M = -0.4\%$, $p > 0.05$) as it increased slightly thereafter. Only the models with exponential hysteresis (2E1 and 2CE1) could match the shapes of the action-history curves, and the addition of constant bias made the correspondence even more precise. Concerning its pitfall of mimicry, constant bias alone (2C) manifests as an across-trial increase in the probability of repetition that superficially resembles the multitrial signature of an extended hysteresis trace.

To better interpret the preceding model-independent time courses, the fitted parameters of the GRL model with either exponential or *n*-back (i.e., 4-back) hysteresis provide context by explicitly factoring out confounds in constant bias as well as the effects of value-based learning (Fig 13). (The selection of 4-back is only for comparison of action-history curves, as the corrected fit of the 9-parameter 2CN4 model was actually worse than that of 2CN2 after adding two more free parameters.) This juxtaposition of parametric and nonparametric implementations of hysteresis revealed notably close correspondence for at least the first two trials back. However subtle the correspondence may be for decaying 3-back and 4-back effects, the superior overall fit of the exponential model relative to a simpler 2-back model (2CN2) already indicated the persistence of collectively significant cumulative effects from 3-back and beyond. Moreover, omission of constant bias (2E1 or 2N4) consistently inflated all of the modeled repetition weights, revealing the source of the mimicry between constant bias and repetition—

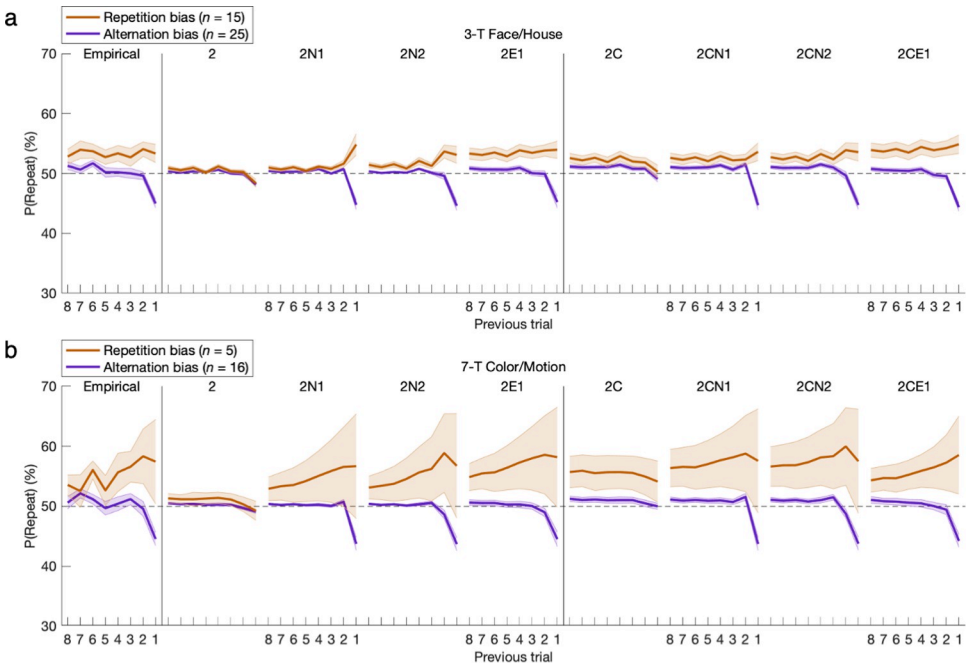

Fig 12. Hysteresis represented across multiple trials. Here the scope of hysteresis was extended to previous actions up to eight trials back. For the repetition-bias group, this probability of repeating a previous action remained elevated above chance prior to 1-back ($p < 0.05$). For the alternation-bias group, this probability instead returned from a 1-back alternation effect ($p < 0.05$) to chance prior to 1-back as it increases backward ($p > 0.05$). Only the models with exponential hysteresis could properly match the shapes of the action-history curves, and the addition of constant bias made the correspondence even more precise. With regard to mimicry, an upward shift in the curve from constant bias in the 2C model superficially resembles the autocorrelational signature of repetition across multiple trials with exponential hysteresis. The nine plots per row each have an identical x-axis despite omission of tick labels from every other plot for readability. Error bars indicate standard errors of the means.

https://doi.org/10.1371/journal.pcbi.1011950.g012

especially in the persistent exponential form—that was alluded to with posterior predictive checks. The 3-parameter adjunct of constant bias and exponential hysteresis proves necessary as well as largely sufficient to distill the action-specific aspects of individual behavioral profiles.

Different forms of bias and hysteresis versus learning performance

The first set of analyses originally split the three levels of learning performance without splitting directions of action biases, whereas the second split directions of bias across Good and Poor learners without splitting levels of learning performance. For this final stage, participants were further divided into six subgroups that separated the two directions of either form of bias as well as the three levels of learning performance—this time also plotting the two directions for previously omitted Nonlearners. There are statistical limitations with this next degree of granularity, which left some of the subgroups with a small sample, but these intersectional subgroups are worth consideration even if only to verify that the main effects essentially extend to this level as well.

With respect to the first set of original findings, action bias and hysteresis were significant for Good learners but even more pronounced for Poor learners and Nonlearners (**Figs 6 and 7**). Second, 2CE1 simulations modeled with constant bias and exponential hysteresis could replicate the directions and magnitudes of empirical action-specific effects both qualitatively and quantitively (**Figs 8 and 9**). Notwithstanding the lack of statistical significance in a few of

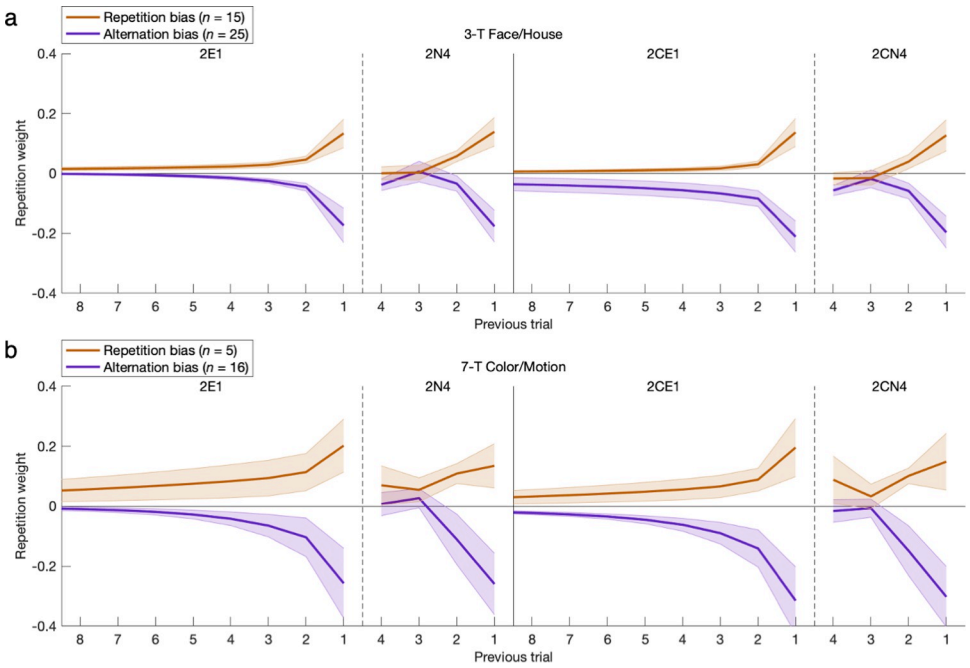

Fig 13. Hysteresis parameters with exponential or nonparametric models. The fitted parameters of the GRL model with either exponential or 4-back hysteresis are plotted as repetition weights (or alternation if negative)—simply β_n for n-back models or the corresponding weights $\beta_1\lambda_H^{n-1}$ in the exponential function. Action-specific effects are better illuminated here by explicitly factoring out effects of RL and GRL within the comprehensive model. There is close correspondence between these parametric (2E1 and 2CE1) and nonparametric (2N4 and 2CN4) implementations of hysteresis for at least the first two trials back. The need for a scope extending beyond 1-back demands more than one free parameter, and a proper hysteresis trace with exponential decay yields an even better fit than a scope of 2-back due to subtle effects from 3-back and beyond. As further evidence of interactions among parameters, omission of constant bias (2E1 or 2N4) consistently inflated the modeled repetition weights as they were forced to attempt to mimic the necessary third parameter for constant bias. Altogether, the CE1 adjunct is essential. Error bars indicate standard errors of the means.

https://doi.org/10.1371/journal.pcbi.1011950.g013

the smallest samples, these trends from either two or three groups consistently held true with the scrutiny of their interface within the six subgroups (**Figs Q and R in S1 Text**).

Alternatives to state-independent action hysteresis

With the primary model comparison establishing that the 2CE1 model has the ideal architecture among the 72 models compared thus far, what follows are other possibilities that could be considered instead of or in addition to state-independent action hysteresis for comparable effects and possible confounds. In other words, these factors could ultimately relate to some form of repetition or alternation across the sequence of action choices. The list of alternative features includes state-dependent action hysteresis $H_t(s_t,a)$ (cf. [21]), state-independent action value $Q_t(a)$, confirmation bias in learning that weighs positive outcomes over negative with the constraint $\alpha_N < \alpha_P$ (i.e., only optimism), or asymmetric learning rates with flexibility in the possibilities for $\alpha_N \neq \alpha_P$ (i.e., optimism or pessimism).

Parsimony is paramount here, and none of these alternatives are as parsimonious as basic hysteresis that is both outcome-independent and state-independent. Take, for example, certain instances of action repetition: Rather than default attribution to a more general optimistic confirmation bias for learning [64–68], first-order perseveration may offer a more parsimonious explanation for some observations. As mentioned for RL, confirmation bias can translate to an asymmetry in learning rates favoring positive over negative outcomes [69–78]—but at the cost

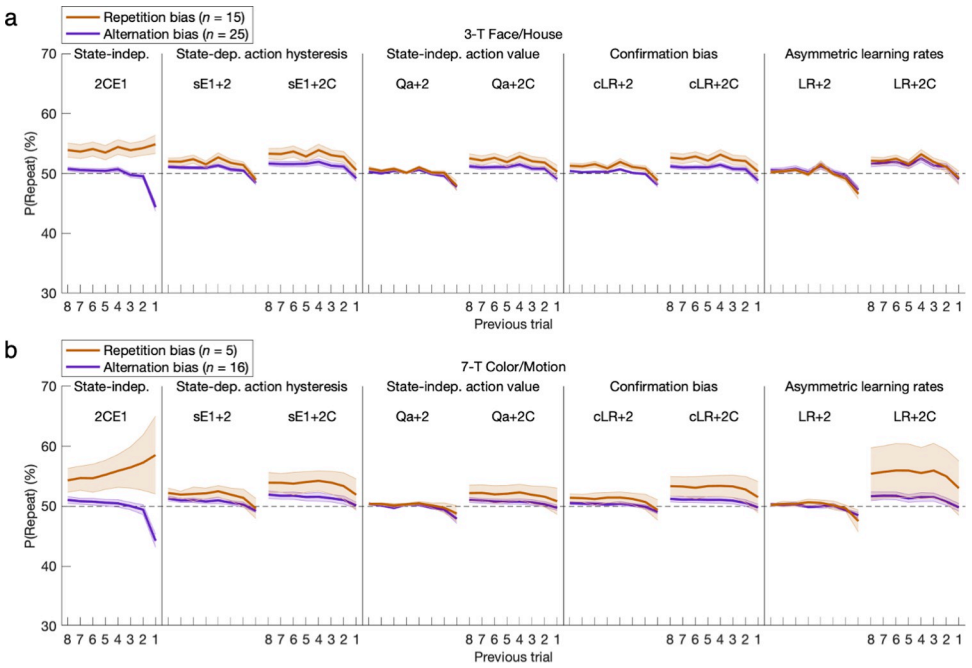

Fig 14. Alternatives to state-independent action hysteresis. Compare to Fig 12. To falsify alternative hypotheses concerning the origins of the apparent effects of state-independent action hysteresis $H_t(a)$ ("2CE1"), the model comparison was first extended to test substitution of state-dependent action hysteresis $H_t(s_t,a)$ ("sE1+2C"), state-independent action value $Q_t(a)$ ("Qa+2C"), confirmation bias in learning with the constraint $\alpha_N < \alpha_P$ ("cLR+2C"), or asymmetric learning rates with no constraint for $\alpha_N \neq \alpha_P$ ("LR+2C"). As expected, none of these alternatives were capable of generating the original action-history curves that only state-independent action hysteresis could produce.

https://doi.org/10.1371/journal.pcbi.1011950.g014

of greater susceptibility to overfitting relative to state-dependent or state-independent hysteresis [42,44,46,79–81], which can manifest its own sort of outcome-independent confirmation bias (see Discussion). (Moreover, as option values become relative in the action policy, the action generalization of GRL can also achieve effects comparable to what asymmetric learning rates might otherwise produce. This point is beyond the present scope but illustrates the broader issue of compounding complexity across the many possibilities that a model could incorporate.)

The initial round of analyses for this extended model comparison began with substitutions of the factors of interest so as to test—and presumably falsify—their alternative hypotheses for the origins of repetition and alternation biases that state-independent hysteresis has been shown to account for with the posterior predictive checks above. Qualitative falsification was indeed robust for all four alternatives, such that none of these model features were capable of generating the original action-history curves that only state-independent action hysteresis could produce (**Fig 14 and Fig S in S1 Text**). These falsifications were hypothesized a priori in consideration of the following conceptual distinctions.

First, state-dependent hysteresis ("sE1+2" or "2sE1") would not align with state-independent hysteresis because the four states were rotated in sequence such that there were variable numbers of trials between the origins and consequences of state-dependent effects. In keeping with this point, only a subtle repetition effect emerged after two trials back. For the original repetition-bias group, the effect sizes were nonexistent for one trial back and quantitatively insufficient from two trials back onward. Furthermore, for the original alternation-bias group, the emergent repetition effect was actually counterproductive such that it pointed in the opposite direction.

Second, state-independent action value ("Qa+2") is unlike state-independent action hysteresis inasmuch as action value is outcome-dependent while action hysteresis is outcome-independent. In principle, there is potential for some degree of confounding if actions that are rewarded consistently end up being repeated consistently. However, in this controlled environment, state-independent action value had little impact on the action-history curves. For the second data set at least, there was a subtle alternation effect in both the original alternation-bias group and the original repetition-bias group—counterproductively for the latter.

Third, confirmation bias in learning ("cLR+2") is generally limited to action repetition and is not only outcome-dependent but also state-dependent in the presence of rotating states here. Like with state-dependent hysteresis, there was only a subtle repetition effect from two trials back onward. However, unlike with state-dependent hysteresis, model simulations for the alternation-bias group did not exhibit a contrary repetition bias.

Fourth, a more flexible asymmetry in learning rates ("LR+2"), including either an optimistic confirmation bias or a pessimistic doubt bias, is again state- and outcome-dependent in the presence of rotating states here. Notably, not all participants in the repetition-bias group adhered to the rule of $\alpha_N < \alpha_P$ in the absence of the constraint forcing confirmation bias. Hence the action-history curve for the repetition-bias group was not elevated above chance beyond 2-back as before with the constrained "cLR+2" result. Instead, the unconstrained asymmetry of "LR+2" produced a 1-back alternation effect for both groups—that is, also counterproductively for the repetition-bias group. With respect to the alternation-bias group, the model's effect was insufficient in magnitude to quantitatively account for the actual effect observed.

Extended model comparison

At this stage, each of the four alternatives had been falsified against state-independent hysteresis with its parsimonious account of the origin of the repetition and alternation effects of interest. The next issue to investigate was the extent to which an alternative feature might instead complement state-independent hysteresis for an even more complex model. Accordingly, the extended model comparison not only substituted these features—namely, state-dependent action hysteresis, state-independent action value, confirmation bias, and asymmetric learning rates—but also added them while crossing with constant bias and 1-back, 2-back, or exponential state-independent hysteresis (e.g., "sE1+2C", "sE1+2CN1", "sE1+2CN2", "sE1+2CE1") in subsets of eight models per alternative (**Table 4 and Figs S-W and Tables Q-U in S1 Text**). The eight models crossed with each alternative feature mirrored the previous reduction of the primary model comparison.

The extended model comparison was applied both within and across the six subsets of eight models (with 44 models in total for the omnibus comparison). The first two subsets built up to constant bias and exponential hysteresis but distinguished the original subset with state-independent hysteresis (e.g., "2CE1") from a new subset with state-dependent hysteresis (e.g., "2CsE1"). The remaining four subsets added each of the four alternative features as a fixed component crossed with the original subset of eight models building up to 2CE1 (e.g., "sE1+2CE1", "Qa+2CE1", "cLR+2CE1", "LR+2CE1").

Within every one of the model subsets, the group-level fitting results consistently favored the addition of the CE1 adjunct with all three of its parameters. In other words, the effects of state-independent hysteresis are indeed substantial, and these specific effects are not confounded with those of any of the alternative features because no alternative could eliminate the need for including state-independent hysteresis in order to adequately fit even the Good-learner groups.

Table 4. Extended model comparison. Additional models were constructed with substitution or addition of the alternative features that might be expected to interact with effects of state-independent action hysteresis. Each alternative was fixed within a new subset of eight models building up to constant bias and exponential state-independent hysteresis ("-CE1"). Variations on substitution of state-dependent hysteresis in particular were also tested up to two parameters. Listed for each participant group are the best-fitting models (per AICc score) among each subset of eight models as well as the full set of 44 models. Although there appears to be some quantitative evidence suggesting state-dependent hysteresis in addition to state-independent hysteresis, the lack of qualitative validation with falsification leaves this quantitative result inconclusive. Hence the 2CE1 model remains preferred for a final model. "df" stands for degrees of freedom. See also Figs S-W and Tables Q-U in S1 Text.

Model comparison	df	Best fit	df	AICc
3FH: Good learner (n = 31)		sE1+2CE1	9	71.13
State-independent action hysteresis	7	2CE1	7	67.69
State-dependent action hysteresis	7	2CsE1	7	67.47
State-indep. + State-dep. action hysteresis	9	sE1+2CE1	9	71.13
State-indep. hysteresis + State-indep. action value	9	Qa+2CE1	9	64.92
State-indep. hysteresis + Confirmation bias	8	cLR+2CE1	8	68.32
State-indep. hysteresis + Asymmetric learning rates	8	LR+2CE1	8	69.57
3FH: Poor learner (n = 9)		sE1+2CN1	8	65.50
State-independent action hysteresis	7	2CE1	7	54.06
State-dependent action hysteresis	7	2CsE1	7	62.40
State-indep. + State-dep. action hysteresis	9	sE1+2CN1	8	65.50
State-indep. hysteresis + State-indep. action value	9	Qa+2CE1	9	50.32
State-indep. hysteresis + Confirmation bias	8	cLR+2CE1	8	54.67
State-indep. hysteresis + Asymmetric learning rates	8	LR+2CE1	8	54.73
3FH: Nonlearner (n = 7)		2CE1	7	16.04
State-independent action hysteresis	7	2CE1	7	16.04
State-dependent action hysteresis	7	2CsE1	7	5.94
State-indep. + State-dep. action hysteresis	9	sE1+2CE1	9	15.96
State-indep. hysteresis + State-indep. action value	9	Qa+2CE1	9	14.82
State-indep. hysteresis + Confirmation bias	8	cLR+2CE1	8	14.43
State-indep. hysteresis + Asymmetric learning rates	8	LR+2CE1	8	14.72
7CM: Good learner (n = 16)		sE1+2CE1	9	72.64
State-independent action hysteresis	7	2CE1	7	67.62
State-dependent action hysteresis	7	2CsE1	7	69.93
State-indep. + State-dep. action hysteresis	9	sE1+2CE1	9	72.64
State-indep. hysteresis + State-indep. action value	9	Qa+2CE1	9	66.07
State-indep. hysteresis + Confirmation bias	8	cLR+2CE1	8	67.07
State-indep. hysteresis + Asymmetric learning rates	8	LR+2CE1	8	67.55
7CM: Poor learner (n = 5)		Qa+2CN2	9	50.02
State-independent action hysteresis	7	2CN2	7	48.15
State-dependent action hysteresis	7	2CsE1	7	23.02
State-indep. + State-dep. action hysteresis	9	sE1+2CN2	9	49.30
State-indep. hysteresis + State-indep. action value	9	Qa+2CN2	9	50.02
State-indep. hysteresis + Confirmation bias	8	cLR+2CN2	8	46.05
State-indep. hysteresis + Asymmetric learning rates	8	LR+2CN2	8	46.78

https://doi.org/10.1371/journal.pcbi.1011950.t004

Next comparing all 44 models across the six subsets at once, there actually was a notable improvement in overall quantitative fit with the addition of state-dependent hysteresis in particular (FH-G: sE1+2CE1, FH-P: sE1+2CN1, FH-N: 2CE1, CM-G: sE1+2CE1, CM-P: Qa+2CN2).

Thus, among the four candidates, state-dependent hysteresis could merit highest priority as the next feature to explore as a possibility for an even larger 9-parameter model. However, despite quantitative gains for state-dependent hysteresis as well as other alternatives, there were still no qualitative improvements in any model-specific effect as would be necessary to falsify a base model having only state-independent hysteresis (**Figs T-W in S1 Text**).

With respect to the otherwise best candidate of state-dependent hysteresis, the absence of qualitative falsification means that its quantitative improvement in fit might actually reflect a spurious relation with residual nonlinearities in the dynamics of learning processes that, unlike hysteresis, are both state-dependent and outcome-dependent. Inevitably, learning is modeled less than perfectly with the current specification of GRL; to take but one example, there are necessary simplifications of a static rather than dynamic learning rate (cf. [81–93]) as well as static generalization weights [12]. The presently inconclusive evidence for state-dependent hysteresis is nevertheless suggestive of the possibility of qualitative validation in future paradigms designed to address follow-up questions about this and other plausible factors directly. However, the most definitive qualitative evidence here is limited to concluding that the final model remains the parsimonious 2CE1 model prioritizing state-independent hysteresis.

Discussion

Summary

These findings have illuminated action bias and hysteresis in the context of active RL so as to suggest that any such study of sequential behavior would benefit from due consideration of these essential variables. Even for some who learn properly, action-specific effects can be so substantial as to actually outweigh the learning effects under primary focus. The modeling inquired beyond basic RL, but two-dimensional GRL as well as constant bias and state-independent hysteresis (2CE1) could all be validated collectively for both quantitative and qualitative individual differences in highly idiosyncratic human behavior. Simpler alternatives to the 3-parameter CE1 adjunct for bias and hysteresis were systematically falsified with factorial model comparison and posterior predictive checks. Conversely, hysteresis models more complex than the 2-parameter exponential function of the CE1 adjunct were susceptible to overfitting. Moreover, an extended model comparison eliminated possible confounds in the form of state-dependent action hysteresis, state-independent action value, confirmation bias in learning, or asymmetric learning rates.

Recognizing each action-bias parameter as fundamental to the core modules of the mixture of experts and nonexperts, the practical costs of these degrees of freedom do not preclude parallel development of learning algorithms and theory. On the contrary, accounting for bias and hysteresis as sources of variance within and between individuals enhances the interpretability of finite behavioral data, which need to be modeled with the independence of each participant preserved. In environments without the symmetric counterbalancing of the present experiment, the limitations of a model with only learning can be even more substantial from spurious correlations between signatures of learning and nonlearning processes. To the extent that the action-specific aspects of bias and hysteresis would also be even more prominent in tasks with more engaging motor responses, proof of concept in this case of trivial motor demands suggests that these effects on choices and actions are as ubiquitous as they are parsimonious and should always be accounted for as a first priority—even with relevance to efficient artificial intelligence as a feature rather than a bug. While fitting at the level of individuals, building from the foundation of this base model—with at least five free parameters for basic RL (0CE1)—is critical to precisely test for whether and how each individual is learning as but a part of interacting with the environment.

Constant bias and lateral bias

Here, the scope for cognitive modeling of motivated behavior is expanded beyond the abstraction of a disembodied brain. Considering that the motor system is the ultimate interface for the actions to be optimized, even low-level sensorimotor processes can constrain the embodied learner. This special case of a binary, bimanual choice task also translates constant bias to a lateral bias.

Although mostly overlooked as part of models of value-based learning, constant bias has occasionally been reported—with and without laterality [12,14–17,21,91,94–99] as well as between acting and not acting for a go/no-go task [100–105]. Even decision making that is not defined by learning—whether value-based [106] or perceptual [80,86,88–90,92,107–112]—can be affected by such stimulus-independent biases with a less obvious role for bias than would be assumed for skillful action-based decision making where physical aspects of action per se have explicit relevance [113].

The decision cost and action cost implicit in such a bias may reflect more than effector-specific motor bias—for example, not only selecting the left hand but also pressing the left button, engaging the left side of abstractly represented egocentric space, attending to the left hemifield of visual space, or embedding a chosen action within subsequences of left and right actions. Asymmetric costs and biases can be considered at all levels of sensorimotor perception, planning, preparation, and execution. Every participant in this neuroimaging study was right-handed for consistency, such that the coexistence of some leftward biases along with the rightward majority demonstrates the significance of not just handedness [114–118] but also a mixture of different levels of representation for nonexpert control.

Lateral biases, for example, can have diverse origins as well. For this sample of Westernized Americans—who are left-to-right readers, for example—eye-tracking studies have demonstrated that people with this cultural background share a propensity for attending to the left side of a display first [106,119–121]. Even more generally, low-level overrepresentation of the left hemifield has been implicated in tasks as basic as line bisection [122]. These biases are in keeping with the innate right-hemispheric dominance of visuospatial attention in the human brain [123–126]. Yet right-to-left (e.g., Hebrew) readers still learn through experience so as to instead exhibit rightward biases [127–129].

In essence, endogenous and exogenous sensorimotor biases are ubiquitous but not always straightforwardly interpretable beyond net effects reflecting a mixture of factors. For example, a leftward visuospatial bias might be at odds with a rightward motor bias in right-handed individuals performing this visuomotor task. There remains substantial ambiguity concerning the distributions of such biases and the relative influences of personal traits (such as handedness) or environmental factors (such as visuospatial cueing). Nevertheless, the key point established here is the need for flexible and fine-grained modeling of the possibilities for biases at the level of individuals.

Bidirectional hysteretic bias

Maintaining the neutral terminology of "hysteresis" as "repetition" versus "alternation", the model here begins with behavioral phenomenology before elaborating on broad unifying theory. That being said, the theoretical construct most often cited with respect to such hysteresis is perseveration, which describes how past responses are repeated regardless of whether or not it is beneficial to do so according to feedback for a new state of the environment [130–135]. Perseveration is linked with the conceptual umbrella of habit to some extent in not being goal-directed. However, habitual phenomena also tend to be more state-dependent, reward-dependent, time-dependent, and intentional than perseverative phenomena [6,136–144]. The

literature has emphasized repetition over alternation as far back as the classic "law of effect", which postulates repetition of rewarded responses but was also complemented by the "law of exercise" that postulates the repetition of past responses regardless of reward outcomes [64,65,137]. Yet an inverted sort of antiperseveration can also manifest with similarly inflexible tendencies toward rhythmic patterns of alternating responses [145–148].

The present study operationalizes perseveration at two levels: first-order, action-level perseveration for repetition of what an agent just did and second-order, policy-level perseveration for what an agent has been doing—either repetition or alternation depending on the circumstances. First-order perseveration aligns with the conventional usage of the term "perseveration" for action repetition in the context of RL, whereas the second-order perseveration emphasized here is less constrained and can result in action alternation as well for an environment such as the controlled one here. The present paradigm did not actually favor alternation per se but nonetheless facilitated it, such that a reward-maximizing policy would incidentally result in more frequent alternation but without any advantage in reward for arbitrary alternation. The hypothesis of second-order perseveration was apparently confirmed in the majority of participants with alternation biases rather than the default repetition biases more often mentioned in the RL literature. Yet, considered further, net effects in output frequencies can also reflect choice and action biases at different levels of representation.

Relatively low-level properties of the motor system can also contribute to alternation more so than repetition. More nonspecific alternation biases can manifest even in perceptual decision making, including neural correlates localized to motor cortex [147]. Whereas motor priming could favor repetition [149–155], motor fatigue could favor alternation if only for an opportunity to rest an effector and recover energy. The general phenomenon of repetition suppression [156,157] extends to the attenuation of signals in the brain's motor areas—and especially premotor cortex—when actions are repeated [158–160]. Such effects may in part reflect the post-movement rebound of beta-band oscillations [161], which are also perhaps analogous to inhibition of return in sensory systems [162–164]. Tendencies toward alternating can also be apparent in arbitrary free choices made without the feedback of any outcome. Whether in expectation of statistical regularities or merely because of limitations in capacity for short-term memory or cognitive control, counterproductive repetition and alternation biases alike can even persist when a person is explicitly instructed to generate maximally random sequences as simple as mental coin flips [165–175].

Perseveration and action repetition in this context have been related to the functions of dopamine [20,144,176–181] (but see [101,182]) as well as perhaps serotonin [177,183] (but see [101]). The theory here can take into account the roles of dopaminergic systems for not only computations such as the reward-prediction error [184–186] but also motivation, vigor, effort, and skillful execution of movement [187–192].

Multiple expert, semiexpert, and nonexpert controllers

The key dynamic variables in the present model are state- and outcome-dependent action value $Q_t(s_t,a)$ and state- and outcome-independent hysteretic bias $H_t(a)$. Having justified these two fundamental modules as a first priority with constant bias, there are then further possibilities to consider for additions to the mixture of expert and nonexpert controllers. As per the extended model comparison, $H_t(a)$ and $Q_t(s_t,a)$ could in principle be complemented by state-dependent, outcome-independent hysteretic bias $H_t(s_t,a)$ (cf. [21]) or state-independent, outcome-dependent action value $Q_t(a)$. However, taking the qualitatively inconclusive gains in model fit observed here as an example, disentangling nonlinear dynamics for multiple types of learning and hysteresis at different levels of representation is nontrivial in practice.

Regarding $H_t(s_t,a)$ for hysteresis that is outcome-independent but instead conditioned on the current external state, there can be an analogous conceptualization of a choice or action itself as a state-dependent reinforcer (i.e., autoreinforcer) motivating repetition in another positive-feedback loop—or a punisher motivating alternation for exploration. Like $H_t(a)$, its counterpart $H_t(s_t,a)$ can also be modeled with the accumulating hysteresis trace [21]. Along with the alternative of a replacing trace (see Methods), another more constrained implementation of hysteretic accumulation could be based on an action-prediction error (or choice-prediction error) with analogy to the reward-prediction error [40,42–47,96,143,144,178,181]. The action-prediction error has been framed as "value-free", but this label and that of $H_t(s_t,a)$ as "habit strength" (cf. [143]) may fail to represent a more endogenous form of subjective value such as with internal positive feedback for repetition (i.e., autoreinforcement) or negative feedback for alternation. The more neutral and bidirectional label of "hysteresis" is preferred here because "habit" not only overemphasizes repetition but also has more specific connotations of stimulus-response associations that may be more semiexpert than truly nonexpert—translating to biases made inflexibly persistent through reinforcement via the reward-prediction error as well [6,135–141,143,144]. Phenomena in the direction of state-dependent and state-independent repetition alike could also be relatable to choice-induced preference change as a reflection of a type of confirmation bias that resolves cognitive dissonance by disregarding feedback altogether [193–199], producing downstream effects comparable to those of confirmation bias with asymmetric learning rates. As discussed in the Results, there is considerable potential for confounds between $H_t(s_t,a)$ and $Q_t(s_t,a)$ as rewarded actions are appropriately repeated within a state, and likewise for $H_t(s_t,a)$ and $H_t(a)$ if different states have overlap in sequences of actions and outcomes.

Regarding state-independent action value $Q_t(a)$, this construct is conceptually constrained to align with repetition of rewarded actions. The most obvious interpretation conflates actions with low-level motor output—in contrast to the high-level goals of actions directed toward external stimuli [95,97,200–203]—but, under the proper circumstances, there could be cognitive and even strategic aspects to state-independent representations as well for semiexpert control. Sequential action representation under uncertainty can be more abstract than just motor control, such as with action chunking in response to working-memory load [204–206]. Concerning the challenge of adding $Q_t(a)$ to the mixture, a confound with $H_t(a)$ can ensue as rewarded actions are more often chosen. Moreover, a confound with $Q_t(s_t,a)$ can also ensue if actions are rewarded similarly across different states.

Levels of representation for decisions, choices, actions, and hysteresis

In contrast to biases more directly linked to motor representations, more abstract cognitive biases may impact sequential behavior as well. Higher-order choice-level biases—as opposed to action-level—can produce comparable effects of sequential dependence in paradigms where motor output is decoupled from perceptual [163,207–213] or value-based [214–217] decisions that do not require learning (i.e., choice hysteresis as opposed to action hysteresis). Complicating interpretation of choice bias or response bias yet further, effects of response history have been shown to parallel, interact with, and even conflict with effects of stimulus history at lower levels of representation in perceptual decision making [81,89,209,211,213,218–225].

For the phenomenology explored here, questions arise as to the contributions of different levels of representation and their integration in the parallelized modularity of the nervous system—ranging from the most abstract level of option choices to the most concrete level of physical motor output. With respect to constant bias $B(a)$, grounding the observed phenomena in the topology of visuospatial and motor representations is more immediately obvious because

intrinsic action cost naturally corresponds to a bias that is both state-independent and sequence-independent. Hence an initial hypothesis here was that rightward biases would be more common among the exclusively right-handed participants, for example.

Whereas constant bias is more straightforward, the origins of even the basic hysteresis emphasized here are more nuanced. Yet that argument also primarily, albeit not exclusively, points to action-based representations—unlike with choice hysteresis as opposed to action hysteresis. First, there is the distinction between state-independent hysteresis $H_t(a)$ and state-dependent hysteresis $H_t(s_t,a)$, which have crucial differences between them despite both being outcome-independent. Whereas state-independent hysteresis may be primarily action-based, this may be less the case for state-dependent hysteresis.

As states of the task environment were rotating while the binary set of actions remained fixed (and time pressure was imposed), a state-independent action representation with tangible visuospatial and motor mapping is unlikely to entail as much abstract representation in terms of a high-level choice rather than action planning and execution. That is, the task incentivizes immediately mapping decisions directly to the space of actions and affordances [226–229], incurring no cost in doing so as long as the motor component of the task is simple and predictable.

In contrast, a state-dependent action representation would more plausibly invoke abstract choice representation to a substantial degree. Insofar as abstraction can be inherent to learning to map an action to the context of an arbitrary state with this sort of instrumental (or operant) conditioning [136,137], a state-aware controller would be making more of an abstract choice about the action than a state-blind controller would. Thus, state-dependent hysteresis could be less contained within action space and instead entail more abstract representation in choice space.

For other situations in which actions might not be as tangible and well-defined as they are in the present setting, greater degrees of abstraction away from action space and into choice space can become more plausible even for state-independent choice hysteresis. Further investigation will be needed for task demands across the spectrum ranging from the present extreme—that of the simplest one-to-one binary mapping across choices and actions as well as effectors and spatial locations—to the opposite extreme of a symbolic choice that must be made either in the absence of any information about subsequent action mapping or in the absence of action altogether (i.e., if only relevant for later actions). Yet the evidence herein is compatible with the majority of active-learning paradigms, where choices typically translate to actions directly and in a straightforward manner.

Dynamics of hysteresis

The specific dynamics of choice or action hysteresis beyond 1-back have typically not been given consideration in previous empirical work with RL and hysteresis for behavior (cf. [79,80,94,95,97,98,101,203,230–242]). Thus far, some computational modeling [12,18,19,21,43,44,46,47,96,181,201,243,244] as well as simpler regression analyses with an autoregressive choice kernel or action kernel [17,20,58–61,245,246] have yielded differing time courses for hysteretic effects, but such findings tend to not be reported in detail.

Following the trends of artificial neural networks, deep learning [247–251], and deep RL [252–260], recent approaches to cognitive modeling have begun to utilize machine learning via the architecture of a recurrent neural network (RNN) [261–263]—such as with a long short-term memory (LSTM) unit [264] or a simpler gated recurrent unit (GRU) [265]—in an attempt to understand core computations for learning (i.e., beyond just nonlinear function approximation for state representation) [266–279]. Whereas such efforts pursue a data-centric approach leveraging predictive power as opposed to the present theory-centric approach

leveraging explanatory power, it is the latter that has so far proven more effective for inference about empirical behavior (but see [87,280]). The mechanistic interpretability of a standard deep-learning approach (cf. [281–292]) is limited with nearly a black box and one typically not amenable to the individual differences here given model demands for data and dimensionality that are orders of magnitude larger. Hence, despite general merits of deep learning, the promise here is confronted by formidable challenges both practical and epistemological. At the very least, deep autoregressive neural networks with inputs for action or choice history—as well as state and reward histories—have begun to speak to not only the degree of nonlinear dynamical complexity but also the significance of sequential hysteresis across longer time scales in parallel with RL [267,270,271,274,277–279].

As part of the motivation for testing different hysteresis traces in the large-scale model comparison here, regression analyses without computational modeling have suggested possibilities for nonmonotonic reversals between short-term alternation and long-term repetition [17,20,58–60] or vice versa [20,61]. Although the dynamics of hysteresis may not always be so complex, sequential patterns can emerge from more than just neural activity persisting from previous trials. On the one hand, amplification of hysteresis over time is possible and can be attributed to working memory and its maintenance of past information [211] or instead to accumulating urgency signals [293] and their baseline activation for a response [294]. On the other hand, phenomena such as the diminishing of hysteresis with longer temporal intervals resonate with an account of sustained residual activity [214,216,295–301]. The exponential function evidenced here is a logical means to monotonic decay and also apt as a matched control against the similarly decaying effects of reinforcement across nonreinforced observations over time [12,18,19,21,43,47,96,243,244].

The primacy of bias and hysteresis as well as individual differences

That the effects illuminated herein are so parsimonious and demonstrably extractable means that comparable studies of RL and other sequential tasks generally stand to benefit from considering bias and hysteresis as part of due diligence—even if the main focus of inquiry is directed elsewhere. Being more representative of actual behavior, the expanded 5-parameter base model 0CE1 aims to enhance parameter identifiability with respect to actual RL as opposed to action-specific components of variance that may mimic or otherwise obscure signatures of learning with spurious correlations [17,18,27,28,39–47]. Before making additional assumptions, parsimoniously imposing action-specific parameters with first priority can be beneficial as a sort of regularization for learning parameters that in practice are nontrivial to extract and estimate.

The present solution of a more comprehensive yet parsimonious model avoids compromising the independence of separate data sets, making it preferable to alternative small-data solutions finding recourse in regularization via fully group-level estimation (i.e., concatenating data sets or averaging parameters) or the intermediate approaches of empirical priors and hierarchical Bayesian modeling across participants [13,29,79,302–305]. From an idealized Bayesian-statistical perspective, compromising independence between individuals in this way mitigates putative measurement error from limited data. From a realistic perspective, however, measurement error and test-retest reliability are irrelevant and ill-defined here: A session of an experiment for a person and their internal state at the moment is a unique, nonrepeatable event—especially for dynamic learning, where model parameters are guaranteed to change over long timespans [47,105,306–320]. Across time, both learning and nonlearning modes for behavior can evolve or discretely alternate with dynamics that are as enigmatic as they are idiosyncratic [81,86,88–93]. In any case, anything resembling measurement error in behavior that

is fitted with an incomplete model is not necessarily more substantial than modeling error [32], including that from omitted variables such as action bias and hysteresis.

As per the bias-variance tradeoff for the nonconvex optimization problem of model fitting, a reduction of variance in parameter fits with the group-level constraints of hierarchical Bayesian estimation necessarily incurs undesirable estimation bias both toward averages across individuals (i.e., shrinkage) and toward the specifications of parametric probability distributions [30–35,321,322]. Whereas a biased estimator will be guaranteed to show greater stability than an unbiased estimator, this property becomes disadvantageous when the biased estimator is less veridical. In a multidimensional parameter space, this estimation bias is exacerbated and can not only underestimate but also overestimate individual differences along a given dimension as a result of complex interactions among parameters constrained by outside data—for example, mimicry of a more constrained parameter by a less constrained one.

There is a more general epistemological problem with inference predicated on the strong assumptions of model validity and a common distribution for every individual from a random grouping of independent data sets, thereby speciously invoking the ecological fallacy [36–38]. The ecological (or population) fallacy is characterized by the principle that, even if a group in the aggregate is representative of the majority of the individuals within said group, any given individual or subgroup is not necessarily representative of the group at all. Hence, when assumed for the individual, assumptions based on group-level or hierarchical inference are inherently fallacious and invalidate potential conclusions about individual differences, including those applied in computational psychiatry and neurology [323–325] for computational phenotyping [29,316,318,319,326–328]. This point is missed in a cognitive-modeling literature now widely and unquestioningly adopting hierarchical Bayesian fitting—a trend motivated by the allure of results that, being biased, merely appear to be cleaner because of unverifiable assumptions about the unknowns of diverse brain states.

With independence instead preserved for each participant, the power of individual differences in computational modeling includes the means to model-based classification of individuals for hypothesis testing within, between, or across subgroups defined qualitatively and quantitatively by various dimensions of a model validated with posterior predictive checks [12,21]. Furthermore, if participants are grouped in advance—as with clinical studies, for example—this approach can address the initial classification in relation to model-based classification as well as model-based metrics across a continuum. More precise individual-level interpretability also extends to model-based analysis of neurophysiological data [329–331], including better estimation of computational signal dynamics within and between individual brains [12,21].

The optimality of nonexpert control with lessons for ML and AI

From an apparently intuitive perspective, any bias or hysteresis in general might be viewed as interference that needs to be mitigated for optimal reward maximization with expert control. Perseveration in particular has a legacy of association with pathologized traits of compulsive behavior, brain lesions, and neurological disorders [20,97,130–132,233,332]. In a somewhat similar vein for the present study, the learners who performed best were not unbiased in this regard but did characteristically exhibit the least bias. Likewise, in experiments with extended training, the relative weight of choice biases tends to decline as learning performance improves over time [333,334]. Both repetition and alternation biases tend to be most robust when evidence is uncertain, confidence is low, and difficulty is high [207,224,238,333,335,336]. Among these factors, that of difficulty is most directly accounted for by the present model with point estimates for action values because these value estimates are rescaled by the nonlinearity of the

softmax policy. That is, bias has greatest impact in the most locally linear vicinity of the intercept of the sigmoid psychometric curve as a function of value difference.

From another perspective, however, nonexpert biases are not suboptimal as part of a tradeoff for optimizing in favor of minimal cost, including computational costs of cognitive demands and motor control as well as sheer time. If uncertainty, unfamiliarity, or irrelevance trivialize a given decision, then choosing quickly according to low-cost biases by default would be optimal to mitigate energy expenditure and fatigue—even if fast responses could not affect the reward rate. Although additional complexities of dynamical decision making [293,337–341] are presently abstracted away for tractability, a speed-accuracy tradeoff [342,343] was evident both within and between participants in the data sets here [12]: More difficult decisions were slower, and across individuals, decisions made by better learners were slower as well. Altogether, the effortful aspects of task engagement can be integrated into the common currency of the cost of control [344–356]. Internal cost-benefit analysis also weighs these costs against reward incentives to determine the level of motivation to effortfully leverage expertise rather than defer to more efficient nonexpert control. Aside from the uncertainty in learning, the monetary incentivization in an experiment tends to be low in subjective value and can be reflected in low levels of motivation and arousal as well as effort and attention.

In contrast to its associations with suboptimality, perseveration has also been framed as adaptive policy compression amid a tradeoff between maximizing expected reward and minimizing the information-theoretic complexity of an action policy [135,176,179,205,206,357–359]. This principle can be extended to higher-order perseveration as well as action or choice bias in general. The dimensions exemplified here reflect how a more state- and outcome-dependent policy trades off being more rewarding for being more complex than a more state- and outcome-independent policy. In addition to undirected exploration with bias rather than variance (i.e., policy stochasticity for the latter), even exploitation can be achieved both more efficiently and more effectively with choice bias as a semi-optimal heuristic for strategic satisficing [360–362] if appropriate for a given environment [81,277,363–365]. In other words, nonexpert biases can even be leveraged in a semiexpert fashion. Such a reward-compressibility tradeoff may offer an analogy with other biases of perceptual stability [208,213,221,223,366,367] or cognitive anchoring [368,369]: Both similarly leverage heuristics for efficiency—whether at the expense of veridical sensory representation or at the expense of precise statistical estimation.

Even low-level motor biases, which if disregarding their benefits in lower internal cost might otherwise be considered a disadvantage of embodiment, may also not be so disruptive as part of a tradeoff for which an embodied RL policy has greater potential for robustness in learning per se. Indeed, embodied RL for concrete actions can achieve greater fluency than disembodied RL for symbolic choices abstracted away from motor output [95,99,202]. Benefits of embodied learning may be facilitated by lesser working-memory demands and lesser overall demands from the topology of the action space as a cognitive map [370–373] more amenable to spatial and embodied representations in the neural circuitry of the basal ganglia and cortex [8,21,22,374–378].

In addition to endogenous choice and action biases, exogenous factors can also shape biases over time. For example, the environment here was structured to be conducive to an alternation bias via second-order perseveration. Adaptive bias has been suggested for actions, effectors, or spatial locations in experimental paradigms delivering rewards asymmetrically with distributions that are congruent or incongruent with respect to particular biases [99,106,111,210,212,213,300,301,333,379]. Adaptive control with the heuristics of a mixture policy would entail flexible leveraging or suppressing of action bias and hysteresis to strike a balance among various tradeoffs of bias and variance, speed and accuracy, energy and effort, benefit and cost, reward and compressibility, expertise and efficiency, or exploration and exploitation.

With analogies between animal learning [6–8,55–57,380–382] and machine learning [49–54,288,291,383–396], the theory of a mixture of experts is based on advantages of modular parallelism and conditional computation for balancing versatility and efficiency in optimal control. As with the mixture-of-experts (MoE) architecture per se (which has also proven effective for sparse scaling of a deep neural network), the scope of this consilient theory can be extended to systems of varying levels of expertise as well as nonexpert controllers and their numerous choice and action biases (cf. [12,21,81,86,88–92,94–96,98,99,143,144]). Benefiting from distributed control of decisions and actions across diverse levels of representation in the networks of the nervous system [227–229], a mixture of experts and nonexperts can dynamically mediate distinct subpolicies with the metacontrol of a manager for arbitration over the gated ensemble of modular learning and nonlearning processes. With adaptive computation for a given subpolicy, semiexpert or nonexpert controllers could be upweighted for conserving time and energy when incentivized, whereas expert learning algorithms could be downweighted for being evaluated as too costly to compute or insufficiently reliable for lack of information or fidelity at any given moment.

Reverse engineering such manifestations of the implicit wisdom of evolution yields a wellspring of inspiration for designing artificial intelligence. Although this computational modeling has primarily been tailored to human behavior and its neural substrates, the fundamental concepts are well-suited for interdisciplinary triangulation across the consilience of RL. With respect to an embodied robotic system, cost and reliability can be factored in for the state of the plant with its physical constraints in action sequences as well as demands for inference and decisions with minimal latency [256,258,260,397–404]—all with analogy to a nervous system characterized by not only metabolic constraints and memory constraints but also motor constraints and embodied cognition [226–229,405,406]. More generally, these insights extend well beyond robotics into all of control theory, machine learning, and artificial intelligence. The costs of time, energy, and computational resources are not limited to active RL and indeed can be found in any system for inference or control. Considering their ubiquity, variants of bias and hysteresis of any abstraction are essential to multiobjective optimization in a resource-limited but resourceful agent—one who is effectively a mixture of agents and at that a mixture of experts and nonexperts.

Methods

Ethics statement

Including functional MRI (fMRI), participants provided informed written consent according to protocols approved by the Institutional Review Board of each of six scanning sites—namely, the California Institute of Technology; Columbia University; New York University; the University of Pennsylvania; the University of California, Santa Barbara; and the University of Southern California.

Preface

In this second report, only the details most relevant for the present purposes are included here. Additional details of the study, including neuroimaging, can be found in the original report for these data sets [12]. Incidentally, "3 T" and "7 T" refer to field strength for the respective MRI scanners.

Participants

Forty-seven (male:female = 27:20; age: M = 25.5 y, SD = 4.9 y) and twenty-two (male:female = 12:10; age: M = 28.0 y, SD = 6.0 y) human participants volunteered for the 3-T Face/House

and 7-T Color/Motion versions of the study, respectively. The 3-T Face/House version was itself multisite, being conducted at five separate facilities for magnetic-resonance imaging (MRI) where participants were recruited from the respective universities and local communities of each laboratory. All participants were screened for MRI contraindications; all were right-handed and generally healthy adults between 18 and 43 years old. Participants in the 7-T Color/Motion version were also screened for color blindness. Upon completing the study, participants were paid $10 for minimizing head movement plus the amount of money earned within the task as the main incentive.

Experimental procedures

A hierarchical reversal-learning task [12] delivered probabilistic outcomes for combinations of categorized states and conditional actions with reward distributions changing across 12 blocks of trials. Note that **Fig 1A** (showing only one state category) does not actually represent a possible sequence of trials (see **Figs A and B in S1 Text**) because the purpose of the figure is instead to conceptually illustrate action bias and hysteresis. To represent each active state (a two-armed contextual bandit), four new cues were assigned randomly every run with two pairs of images each respectively drawn from two state categories. In the version of the experiment incidentally conducted with a 3-T MRI scanner, these categories were faces and houses (images in Fig 1 courtesy of [407]).

At the onset of each episodic (i.e., separate) trial, one of four predictive cues was presented with equal probability, but trials were also ordered in a series of randomized and counterbalanced quartets that each included four cues representing separate states. These quartets were constrained such that a cue never appeared in consecutive trials. The onset of a trial was marked by a face or house image appearing. The participant was allotted 2 s to respond to this active state by pressing one of two buttons with the corresponding index finger of either the left or right hand. A fixed interstimulus interval (ISI) of 3 s separated the cue and the outcome.

The transition probabilities for the action given the state determined whether the outcome following the ISI was a rewarded state or a nonrewarded state. Delivery of an actual reward of $0.30 was symbolized by an image of a dollar sign for 1 s, whereas a scrambled dollar sign signified an absence of monetary reward for that trial. The duration of the jittered intertrial interval (ITI) was drawn without replacement within a run from a discrete uniform distribution ranging from 3 to 7 s in increments of 41.7 ms. If the participant failed to respond in time, the nonrewarded outcome appeared immediately as the fixation cross turned red for 1 s; the ISI would then be merged with the subsequent ITI.

Twelve blocks of trials were defined by permutations of three experimental conditions, The first condition for category value had three possibilities also counterbalanced within a run. This condition determined whether the face category had greater, lesser, or equivalent value relative to the house category. For the unequal conditions, the category with greater value included reward probabilities of 62.5% and 100%, whereas the category with lesser value included reward probabilities of only 43.75%. For the equal condition, both categories included reward probabilities of 43.75% and 81.25%. These exact probabilities were all divisible by sixteenths and so were evenly split between two 32-trial blocks with 8 trials per state. (For the odd probabilities of 43.75% and 81.25%, the more-rewarded halves of the distributions were evenly distributed within a condition sampled across runs: The net probability of 43.75% (7/16) was the average of 37.5% (6/16) and 50% (8/16), and net 81.25% (13/16) was the average of 75% (12/16) and 87.5% (14/16).) A nonzero reward probability was only assigned to one action per state, always leaving an alternative action with zero probability of reward. This complementarity between actions within a state was designed to reveal action generalization.

The second condition for state value had two possibilities partially counterbalanced with a 2:1 ratio within a run. This condition concerned which state (arbitrarily "A" or "B") had the greater value within a category if the category included two different reward probabilities for a given block.

The third condition for action mapping had four possibilities. This condition concerned the mapping of a state category's reward probabilities to actions, such that the two states ("A" and "B") within a category always symmetrically provided rewards for opposite actions. The possibilities for this condition could be summarized across all four active states like so: "LR&LR", "LR&RL", "RL&LR", or "RL&RL", where the example of "LR&RL" can be expanded as "AL/BR & AR/BL" for the binary hierarchical metastates of the face and house categories, respectively. That is, "LR&RL" (or "AL/BR & AR/BL") would mean that the left action is rewarded for face A and house B while the right action is rewarded for face B and house A. This complementarity between states within a category was designed to reveal state generalization.

Rather than sheer randomness in the design, which would especially limit interpretation of individual differences, meticulously controlled counterbalancing was crucial for eliminating confounds within and across individual sessions. For each participant, different conditions were randomized and counterbalanced to evenly distribute rewards for categories, states, and actions in a factorial design defining 12 blocks that included hierarchical reversals of instrumental learning. Four scanning runs including three blocks each and 32 trials per block amounted to 384 trials in total. (Prior to the actual experiment, the participant completed 10-trial practice sessions with separate stimuli both outside and inside the scanner.)

Nearly attaining a 3 x 2 x 4 design ("category value" x "state value" x "action mapping") for the 12 blocks, the 3 x 2 and 3 x 4 crosses were fully counterbalanced while the 2 x 4 cross could only be partially balanced given the number of blocks. By virtue of this counterbalancing, choosing the same action for every single trial of the session was guaranteed to yield exactly half of the available rewards. Likewise, each state category preceded exactly half of the available rewards within each run. Moreover, with reward probabilities in units of sixteenths, each run included exactly or nearly one quarter of the rewards for the entire session. Yet the reward probabilities for state-action pairs fluctuated from block to block so as to facilitate variability in the dynamics of neural signals of interest. Across the session, what remained constant amid these fluctuations was the anticorrelational pattern between complementary actions within a state and between complementary states within a category. The categories were independent of each other without any such structured pattern between them.

Between blocks, the design was constrained for a single remapping to mark the onset of a new block within a run, where reversals of rewarded actions occurred for only one category at a time. The two categories were remapped in turn in a random order counterbalanced across runs, such that each category had exactly one between-block remapping per run. Although the participant was informed that the reward probabilities could change throughout the session, no explicit indications were provided as to how or when such changes might occur.

Regarding the 7-T Color/Motion version conducted in parallel, this second version of the experiment was mostly matched to the first but was not entirely identical. The main difference was that the 7-T version substituted dynamic colors and directions of motion in lieu of faces and houses as state categories. Moreover, these color and motion stimuli (4 in total) were not replaced every run as with the 3-T version's faces and houses (16 in total). Although the two pairs of visual stimuli comprising the two categories instead remained constant across the entire session, the counterbalanced factorial design of the 3-T version was preserved such that the reward probabilities for the respective states still rotated as before.

Computational modeling: Generalized reinforcement learning

Generalized reinforcement learning (GRL) [12] is a quasi-model-based extension of model-free reinforcement learning (RL) [1–3]. The description of GRL that follows here is simplified so as to shift the emphasis to details of action-specific bias and hysteresis in the model's mixture policy for action selection (**Fig 1**). Importantly for the present purposes, GRL adds even more complexity to the mixture of experts and nonexperts. Incidentally, this complexity takes the form of intersecting dichotomies for associative versus discriminative generalization and state versus action generalization. This expansion of RL in parallel with the expansion of action bias and hysteresis serves to demonstrate the practical feasibility of simultaneously investigating more complex learning theory despite the costly degrees of freedom inherent to the added complexities of the nonlearning modules.

Neuroimaging analysis [329] and thus the original critic/Q-learner (CQ) model [12,21] are presently set aside for this analysis of the single-step cue-outcome task. This simplified version of the GRL model omits not only passive state-value learning—which would be via the critic module of the actor/critic architecture [408–410]—but also the temporal-difference (TD) prediction method [411–413]. Given the absence of the TD update here, the action-value learning that remains also makes no distinction between off-policy and on-policy methods such as in the Q-learning algorithm [414,415] and the state-action-reward-state-action (SARSA) algorithm [416], respectively.

The beginning of a run marks initialization of action values $Q_t(s,a)$ for all novel state-action pairs. As representing priors in the absence of previous associations would entail some kind of internal model, a naïve model-free agent initializes to zero [417]:

$$\forall(s,a) : Q_0(s,a) = 0$$

The rotating active states were initiated with the onset of each trial. Upon transitioning from an active state to an outcome state, a reward-prediction error (RPE) δ_{t+1} is determined by the discrepancy between the current action-value estimate $Q_t(s_t,a_t)$ and the subsequent reward (or lack thereof) r_{t+1} presented in the binary outcome state. The RPE would obey the same equation with any scalar reward as well:

$$\delta_{t+1} = r_{t+1} - Q_t(s_t,a_t)$$

As with any standard RL model, the value of the chosen state-action pair is updated according to the following delta-learning rule with a fitted learning rate α (for $0 \leq \alpha \leq 1$):

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha\delta_{t+1}$$

The equations thus far have described the basic RL model in its original form. In preparation for the following section on GRL, note again that the reward magnitude is fixed for this paradigm. Hence the cached action value $Q_t(s,a)$ effectively corresponds to the estimated probability of reward. To prevent the duplicated and relayed prediction errors of GRL from producing an illogical expected value for probabilistic binary outcomes (i.e., $0 \leq P \leq 1$), the clipping function $f(x)$ clips action value between zero and unity as an ad-hoc solution for this particular case where subjective value represents probability. Although reference dependence and normalization are mechanisms of relevance to value-based learning [418–421], the present paradigm is not suitably amenable to these complexities. Possibilities for alternatives to clipping are not considered for now inasmuch as a guaranteed improvement in fit in the absence of this constraint would presently be uninterpretable: Probability estimates above unity or below zero would be meaningless as probabilities per se, and a negative value would also correspond to negative valence despite an absence of punishment. When this neural model is

applied to (computational) model-based neuroimaging analysis [12], these simulated signals have substantial implications for the interpretation of value signals in the brain, which would be maximized with certain reward and range from neutral to appetitive rather than including anything in the aversive range of valence. The x here refers to an updated value estimate prior to transformation:

$$f(x) = \text{clip}\{x, [0, 1]\}$$

In contrast to previous RL models, the GRL model introduced here additionally applies a common RPE signal to learning about other state-action pairs within the same state as well as the same state category. Aside from generalization, the value of any state-action pair not encountered remains as is rather than being subject to decay or "forgetting" with potential for overfitting [43,58,200,422–426]. (For future investigation elsewhere, there are intriguing parallels to note in the mathematics of value decay versus counterfactual updating for non-encountered representations.) Presently, the two-alternative forced choice allows for a straightforward model of discriminative action generalization, such that the nonchosen action a'_t receives an inverse value update as the complement of the chosen action a_t (where prime notation refers to complementarity here). The variables a_L and a_R stand for the left and right actions:

$$a'_t = \begin{cases} a_R, & a_t = a_L \\ a_L, & a_t = a_R \end{cases}$$

This counterfactual update is regulated by a negative parameter for the action-generalization weight g_A (for $-1 \leq g_A \leq 0$) that modulates the original learning rate. Although associative action generalization is a possibility elsewhere, this parameter is not allowed to be positive here because the effective input to the softmax function is the difference between two action values—rendering overgeneralization across actions essentially indistinguishable from a mere absence of learning. The constraint that absolute generalization weights do not exceed unity first resolves the potential nonidentifiability issue of multiplied free parameters for generalized delta learning. Conceptually, this constraint also reflects the assumption—one shared with the eligibility trace of the "TD(λ)" algorithm [3,411–413,427,428]—that generalized RPE signals would not be relayed with greater gain than the original RPE signal but rather with lesser or equal gain. (In a different setting, this assumption might be relaxed under the appropriate circumstances.) As with the state generalization that follows, this action generalization is analogous to the temporal generalization of TD(λ) (see [12]):

$$Q_{t+1}(s_t, a'_t) = f(Q_t(s_t, a'_t) + g_A \alpha \delta_{t+1})$$

With only two states per category, state generalization entails an analogous formula where —in addition to the encountered state s_t—the other, complementary within-category state s'_t receives a relayed value update. The variables s_A and s_B refer to state A and state B (arbitrarily designated as such):

$$s'_t = \begin{cases} s_B, & s_t = s_A \\ s_A, & s_t = s_B \end{cases}$$

This update is regulated by a state-generalization weight g_S (for $-1 \leq g_S \leq 1$) that modulates the learning rate. Unlike overgeneralization across actions, overgeneralization across states within a category can be detected here. That is, the agent could incorrectly operate as if the category itself were assumed to be a unitary state ($g_S = 1$), or the agent could at least partially conflate representations of exemplars within a category due to fuzzy boundaries ($0 < g_S < 1$). The

present paradigm is characterized by anticorrelational linkage between states within a category. Hence a negative sign for g_S correctly produces discriminative generalization, while a positive sign for g_S incorrectly produces associative overgeneralization:

$$Q_{t+1}(s'_t, a_t) = f(Q_t(s'_t, a_t) + g_S \alpha \delta_{t+1})$$

The two factors of action generalization and state generalization interact multiplicatively to also update the complementary action for the complementary state. In the optimal case combining discriminative generalization across both dimensions (i.e., $-1 \leq g_A < 0$ and $-1 \leq g_S < 0$), this interactive state-action generalization weight would appropriately be associative ($0 < g_S g_A \leq 1$) for the one state-action pair that is correlated with the original pair rather than anticorrelated:

$$Q_{t+1}(s'_t, a'_t) = f(Q_t(s'_t, a'_t) + g_S g_A \alpha \delta_{t+1})$$

Computational modeling: Mixture policy with bias and hysteresis

The learned Q values are inputs to a probabilistic action-selection policy $\pi_t(s,a)$ characterized by the Boltzmann-Gibbs softmax function and the Shepard-Luce choice rule as a discriminative model of decision making [3,23–25] rather than a generative model. The approximation of a softmax function—effectively with perfect subtraction between two alternatives here—has some limitations in accounting for nonlinearities in actual behavior due to the dynamics of underlying decision processes in the brain [340], but this simplification can suffice for the present purposes as a standard assumption for active-learning models.

In addition to an essential module for action value, the mixture policy here also incorporates inputs from modules for action-specific bias and hysteresis (**Fig 1**) [12,21]. Constant bias $B(a)$ becomes a lateral bias between left and right actions in this case, whereas the dynamic hysteretic bias $H_t(a)$ (cf. [17,18]) maps repetition and alternation to positive and negative signs, respectively. These state- and outcome-independent action biases complemented the state- and outcome-dependent action values to determine the mixture policy's action probabilities via the following softmax function with temperature τ (for $\tau > 0$), which regulates the stochasticity of choices reflecting noise as well as exploration against exploitation [3,429–435]. This policy equation also reduces to a logistic function in the present case of a two-alternative forced choice:

$$\pi_t(s_t, a) = P(a_t = a | s_t) = \frac{\exp\{(Q_t(s_t, a) + H_t(a) + B(a))/\tau\}}{\sum_{a^*} \exp\{(Q_t(s_t, a^*) + H_t(a^*) + B(a^*))/\tau\}}$$

With n-1 parameters for n available actions, constant bias is reduced to a single parameter for a binary action space such as the present one. The indicator function $I_R(a)$ is used for a lateral bias with the arbitrary convention that a positive sign for the parameter β_R corresponds to a rightward bias while a negative sign corresponds to a leftward bias:

$$I_R(a) = \begin{cases} 0, & a = a_L \\ 1, & a = a_R \end{cases}$$

Avoiding the dummy-variable trap, the bias terms are then β_R for the right-hand action and null for the left-hand action:

$$B(a) = \beta_R I_R(a)$$

Modeling action hysteresis in terms of the dynamics of integrated repetition or alternation biases first requires an initialization of the hysteresis trace and its cumulative bias variable $H_t(a)$:

$$\forall a : \; H_0(a) = 0$$

A counter variable C_t is initialized at the beginning of each run to index the total number of actions performed within the run:

$$C_0 = 0$$

This action-counter variable is simply incremented with each action performed:

$$\forall a_t : \; C_t = C_{t-1} + 1$$

Using this action index throughout the run, the indicator function $I_{Ct}(a)$ tracks action history:

$$I_{C_t}(a) = \begin{cases} 0, & a \neq a_t \\ 1, & a = a_t \end{cases}$$

In its currently preferred form ("-E1" models such as 2CE1), the hysteretic bias is determined by its initial (i.e., 1-back) magnitude β_1 and inverse decay rate λ_H (for $0 \leq \lambda_H \leq 1$), where this base of the exponential function is notated as the complement of (i.e., unity minus) the exponential decay rate. A positive magnitude for this autocorrelation ($\beta_1 > 0$) represents a repetition bias in favor of repeating previous actions, whereas a negative magnitude ($\beta_1 < 0$) represents an alternation bias in favor of switching between actions. By conventions with analogy to the eligibility trace of TD(λ) [3], the hysteresis trace (i.e., action kernel) is specified as an accumulating trace rather than a replacing trace so as to not be overly constrained; the latter instead has an upper bound at β_1 and disregards consecutive repeats (cf. [18]). Yet it is ultimately the difference between the cumulative hysteresis effects of competing actions that determines their net weight in the action policy. An accumulating repetition bias ($\beta_1 > 0, \lambda_H > 0$) means that a repeated action would become even more likely to be repeated again with successive repetitions in a positive-feedback loop. Conversely, an accumulating alternation bias ($\beta_1 < 0, \lambda_H > 0$) means that a second repetition would become even less likely. The exponential decay of a given action's bias proceeds indefinitely with each action executed as the hysteresis trace is continually integrated into the cumulative hysteretic bias $H_t(a)$:

$$H_{t+1}(a) = \sum_{i=0}^{C_t - 1} \beta_1 \lambda_H^i I_{C_t - i}(a)$$

The label of the preferred 2CE1 model stands for 2-parameter GRL ("2"), constant bias ("C"), and 1-back exponential hysteresis ("E1")—that is, one degree of freedom preceding exponential decay. This model described thus far includes seven free parameters altogether—namely, learning rate α, action-generalization weight g_A, state-generalization weight g_S, softmax temperature τ, rightward (or leftward) bias β_R, and initial magnitude β_1 coupled with inverse decay rate λ_H for the exponential decay of the repetition (or alternation) bias. An additional 23 models of the 72 in the primary model comparison (**Table 2 and Table A in S1 Text**)

were also nested within the 2CE1 model: X, XC, XN1, XCN1, XE1, XCE1, 0, 0C, 0N1, 0CN1, 0E1, 0CE1, 1, 1C, 1N1, 1CN1, 1E1, 1CE1, 2, 2C, 2N1, 2CN1, and 2E1.

Beyond 1-back hysteresis, the remaining 48 models extended n-back hysteresis with N free parameters β_n for N total previous actions. With reference to statistical fundamentals of generic sequence or time-series modeling, notation with "β" for bias reflects analogous notation for autoregressive and intercept terms corresponding to hysteresis and constant bias, respectively. The signed individual weights β_n each independently correspond to a bias in favor of repetition ($\beta_n > 0$) or alternation ($\beta_n < 0$) of the respective previous action from n actions back. The dynamic hysteretic bias $H_t(a)$ is more generally defined by this flexible equation to accommodate any combination of first n-back and then exponential hysteresis in series —here the first and second terms, respectively, summing backward across time again:

$$H_{t+1}(a) = \sum_{n=1}^{N} \beta_n I_{C_t-n+1}(a) + \sum_{i=N+1}^{C_t} \beta_N \lambda_H^{i-N} I_{C_t-i+1}(a)$$

Computational modeling (extended): Alternatives to state-independent action hysteresis

At this point, the final 2CE1 model has been described in its entirety, and likewise for the other 71 models included in the primary model comparison. What follows are the details of models subsequently tested in an extended model comparison controlling for alternative features that might be expected to interact with the effects of the state-independent action hysteresis presently emphasized (**Table 1**).

Computational modeling (extended): State-dependent action hysteresis

The first alternative feature considered as part of the extended model comparison was state-dependent hysteresis $H_t(s_t,a)$ (cf. [21]) in contrast to state-independent hysteresis $H_t(a)$ as described above. The mathematical specifications of the hysteresis trace are entirely analogous with the incorporation of state dependence.

In this case, the cumulative bias variable is initialized for every state-action pair rather than just actions:

$$\forall (s,a) : \; H_0(s,a) = 0$$

The counter variable becomes a vector $C_t(s_t)$ that instead indexes action counts separately for each state:

$$\forall s : \; C_0(s) = 0$$

This action-counter variable is incremented with each action as before:

$$\forall a_t : \; C_t(s_t) = C_{t-1}(s_t) + 1$$

The indicator function $I_{Ct(s)}(s,a)$ then tracks action history within each state:

$$I_{C_t(s_t)}(s_t,a) = \begin{cases} 0, & a \neq a_t \\ 1, & a = a_t \end{cases}$$

In its pure exponential form ("sE1"), state-dependent hysteresis is determined by its initial (i.e., 1-back) magnitude $\beta^S{}_1$ and inverse decay rate λ_S (for $0 \leq \lambda_S \leq 1$)—now for exponential

decay across only the actions performed within a state:

$$H_{t+1}(s_t, a) = \sum_{i=0}^{C_t(s_t)-1} \beta_1^S \lambda_S^i I_{C_t(s_t)-i}(s_t, a)$$

In addition to the seven free parameters of the 2CE1 model, the extended "sE1+2CE1" model adds two more—that is, $\beta^S{}_1$ and λ_S—for a maximum of nine parameters in total. However, in another subset of models matching the reduced model comparison (2sN1, 2sN2, 2sE1, 2CsN1, 2CsN2, and 2CsE1), state-dependent hysteresis was instead substituted for its state-independent counterpart to remain at most seven free parameters for that subset. The general equation for any combination of first *n*-back and then exponential state-dependent hysteresis is the following:

$$H_{t+1}(s_t, a) = \sum_{n=1}^{N} \beta_n^S I_{C_t(s_t)-n+1}(s_t, a) + \sum_{i=N+1}^{C_t(s_t)} \beta_N^S \lambda_H^{i-N} I_{C_t(s_t)-i+1}(s_t, a)$$

The extended "sE1+2CE1" model thus adds yet another term to the mixture policy:

$$\pi_t(s_t, a) = \frac{\exp\{(Q_t(s_t, a) + H_t(s_t, a) + H_t(a) + B(a))/\tau\}}{\sum_{a^*} \exp\{(Q_t(s_t, a^*) + H_t(s_t, a^*) + H_t(a^*) + B(a^*))/\tau\}}$$

Computational modeling (extended): State-independent action value

In parallel along the dimension of state dependence, the next alternative feature was state-independent action value $Q_t(a)$ in contrast to state-dependent action value $Q_t(s,a)$ as described above. In this case, action values are initialized for not only state-action pairs but also actions per se:

$$\forall a : \; Q_0(a) = 0$$

An action-specific RPE $\delta^A{}_{t+1}$ is determined by the discrepancy between the state-independent action-value estimate $Q_t(a_t)$ and the subsequent reward (or lack thereof) r_{t+1}:

$$\delta_{t+1}^A = r_{t+1} - Q_t(a_t)$$

Naturally, the value update for the chosen action follows an analogous delta-learning rule with an action-specific learning rate α_A (for $0 \leq \alpha_A \leq 1$):

$$Q_{t+1}(a_t) = Q_t(a_t) + \alpha_A \delta_{t+1}^A$$

In addition to the seven free parameters of the 2CE1 model, the extended "Qa+2CE1" model adds two more—that is, action-specific learning rate α_A (for $0 \leq \alpha_A \leq 1$) and action-specific value weight w_A (for $0 \leq w_A \leq 1$)—to reach its maximum of nine parameters. (For the sake of tractability here, action generalization is presently omitted for state-independent action value, but either a shared or tenth parameter could have been added with a generalized RPE updating the state-independent value representation for the nonchosen action.) The weighting parameter between state-independent and state-dependent action value can be incorporated into the mixture policy like so:

$$\pi_t(s_t, a) = \frac{\exp\{(w_A Q_t(a) + (1 - w_A) Q_t(s_t, a) + H_t(a) + B(a))/\tau\}}{\sum_{a^*} \exp\{(w_A Q_t(a^*) + (1 - w_A) Q_t(s_t, a^*) + H_t(a^*) + B(a^*))/\tau\}}$$

Computational modeling (extended): Asymmetric learning rates and confirmation bias

Rather than adding another module to the original mixture policy, another alternative feature that could similarly relate to the repetition or alternations of actions is asymmetry in learning rates between positive and negative RPE signals (α_P and α_N for $0 \leq \alpha_P \leq 1$ and $0 \leq \alpha_N \leq 1$). One subset of eight models ("LR+") flexibly allowed for either an optimistic confirmation bias ($\alpha_N < \alpha_P$) or a pessimistic doubt bias ($\alpha_P < \alpha_N$), whereas another subset of eight models ("cLR+") was constrained with an assumption of only confirmation bias if any asymmetry ($\alpha_N \leq \alpha_P$). Imposing the latter constraint was in keeping with precedent in the modeling literature that emphasizes choice or action repetition by way of optimism and confirmation bias, implying that these forces would ultimately override pessimism and doubt. This modification entailed the addition of only one free parameter for a maximum of eight total in the "LR+2CE1" and "cLR+2CE1" models.

With positive learning rate α_P and negative learning rate α_N, the delta-learning rule is bifurcated with a conditional rule separating positive and negative RPE signals in this new equation:

$$Q_{t+1}(s_t, a_t) = \begin{cases} Q_t(s_t, a_t) + \alpha_P \delta_{t+1}, & \delta_{t+1} > 0 \\ Q_t(s_t, a_t) + \alpha_N \delta_{t+1}, & \delta_{t+1} < 0 \end{cases}$$

For the rewards of fixed magnitude here, the conditions of positive or negative RPE ($\delta_{t+1} > 0$ or $\delta_{t+1} < 0$) would be met in the presence or absence of reward ($r_{t+1} = 1$ or $r_{t+1} = 0$), respectively. Furthermore, with the extension of GRL, these separate learning rates likewise take effect for generalized RPE signals according to analogous conditional updates:

$$Q_{t+1}(s_t, a'_t) = \begin{cases} f(Q_t(s_t, a'_t) + g_A \alpha_P \delta_{t+1}), & \delta_{t+1} > 0 \\ f(Q_t(s_t, a'_t) + g_A \alpha_N \delta_{t+1}), & \delta_{t+1} < 0 \end{cases}$$

$$Q_{t+1}(s'_t, a_t) = \begin{cases} f(Q_t(s'_t, a_t) + g_S \alpha_P \delta_{t+1}), & \delta_{t+1} > 0 \\ f(Q_t(s'_t, a_t) + g_S \alpha_N \delta_{t+1}), & \delta_{t+1} < 0 \end{cases}$$

$$Q_{t+1}(s'_t, a'_t) = \begin{cases} f(Q_t(s'_t, a'_t) + g_S g_A \alpha_P \delta_{t+1}), & \delta_{t+1} > 0 \\ f(Q_t(s'_t, a'_t) + g_S g_A \alpha_N \delta_{t+1}), & \delta_{t+1} < 0 \end{cases}$$

Model fitting and comparison

Whereas the original model comparison permuted models for all variants and reductions of RL and GRL (or fully model-based learning algorithms) [12], the primary model comparison here permuted fewer learning variants to instead combine these with varied implementations of action bias and hysteresis for 72 models in total (**Table 2 and Table A in S1 Text**). Specifically, this model comparison crossed factors for value-based learning, constant bias, *n*-back hysteresis, and exponential hysteresis. The first two factors for learning were limited to the cases of no learning ("X") ($\alpha = g_A = g_S = 0$), basic RL ("0") ($g_A = g_S = 0$), 1-parameter GRL ("1") ($g_A = \min\{0, g_S\}$, $-1 \leq g_S \leq 1$), and 2-parameter GRL ("2") ($-1 \leq g_A \leq 0$, $-1 \leq g_S \leq 1$). (Note that 1-parameter GRL here still refers to two-dimensional GRL but with a shared single parameter.)

With respect to bias and hysteresis, the first main factor was the inclusion ("C") or exclusion of the constant lateral bias β_R, amounting to 36 models each for either possibility. The second

main factor of hysteresis was further subdivided between *n*-back ("N") hysteresis and exponential ("E") hysteresis as nonparametric and parametric alternatives—but not mutually exclusive alternatives—with 40 pure *n*-back models, 8 pure exponential models, and 16 hybrid models. Nonparametric *n*-back hysteresis was tested up to 4 trials back in the presence of learning and up to 8 trials back in the absence of learning. Parametric exponential hysteresis was defined by exponential decay but, when hybridized, allowed up to 2 additional degrees of freedom for nonparametric weights on the most recent previous actions. For example, considering 2-parameter GRL models, *n*-back hysteresis was represented up to 4-back in pure form or 3-back in post-exponential form as 2N2, 2N3, 2N4, 2CN2, 2CN3, 2CN4, 2E2, 2E3, 2CE2, and 2CE3. The 2CN4 and 2CE3 models had the greatest number of free parameters with nine in total.

The competing models were all fitted to empirical behavior via maximum-likelihood estimation with independence maintained at the level of individual participants. Free parameters were optimized for overall goodness of fit to a participant's sequence of actions with randomly seeded iterations of the Nelder-Mead simplex algorithm [436]. All modeling and fitting procedures were programmed with Matlab. The Akaike information criterion with correction for finite sample size (AICc) [62,63] provided a means to adjust for model complexity when comparing models that differ in degrees of freedom. Whereas the XCE1 model with constant bias and exponential hysteresis functioned as the null model for the original model comparison validating GRL [12], here the 0-parameter chance model "X" was used instead for the baseline explanatory power of a completely random action policy. Each free parameter was thus added incrementally with a requirement of statistical justification for every single one.

To further verify the discriminability of the preferred 2CE1 model with its seven free parameters, each fitted instantiation of the model was subsequently used to simulate a data set yoked to that of the respective participant. Another complete model comparison was conducted for these simulated data as a test of model recovery that would indicate whether the 2CE1 model could be discriminated reliably among both simpler and more complex alternatives. Tests of parameter recovery followed with the expectation that the fitted parameters for the simulated data would be correlated with the original fitted parameters for the empirical data that the simulations were derived from. For juxtaposition, these procedures were also repeated with simulations generated by the no-bias model "2" with only GRL.

Following the primary model comparison with its 72 models was the extended model comparison with 44 models spanning six subsets of eight models each. Moreover, each subset of eight models matched the original subset of eight initially highlighted within the primary model comparison—namely, 2, 2N1, 2N2, 2E1, 2C, 2CN1, 2CN2, and 2CE1. The first subset was the original subset itself. The second subset substituted state-dependent hysteresis in six of the original eight models (e.g., "2CsN1", "2CsN2", "2CsE1"). The remaining four subsets added each of the four alternative features—namely, state-dependent action hysteresis, state-independent action value, confirmation bias, and asymmetric learning rates—as a fixed component crossed with the original subset of eight models building up to 2CE1 (e.g., "sE1+2CE1", "Qa+2CE1", "cLR+2CE1", "LR+2CE1"). Comparisons were made both within and across the six subsets.

Data analysis

The group assignments for participants based on learning performance were maintained from the original model comparison [12]. The first measure of performance began with calculating overall accuracy as the proportion of actions for which the participant correctly chose the option that could result in delivery of a reward, excluding choices made for initial encounters with novel cues. Accuracy was compared with the chance level of 50% for each participant using a one-tailed binomial test. A subset of participants was initially set aside as the "Good

learner" group if the accuracy score was significantly greater than chance [18]; subsequent modeling could also confirm that this label was appropriate for each individual within the group. The remaining participants with chance accuracy were subsequently assigned to either the "Poor learner" group or the "Nonlearner" group according to whether or not one of the original learning models could yield a significant improvement in goodness of fit relative to the XCE1 model, which was nested within each learning model while retaining bias and hysteresis but omitting any sensitivity to actual reward outcomes [21].

Individually fitted parameters of the 2CE1 model for action-specific effects were first tested against empirical measures for validation. Omitting the Nonlearner group for additional rigor, correlations were tested for between the rightward bias β_R and the probability of a right-hand action, between the repetition bias β_1 and the probability of a repeated action, and between overall bias $|\beta_R|+|\beta_1|$ and the probability of a correct action (hypothesizing an inverse relation). Linear regression was performed with one-tailed one-sample *t* tests and reported with the Pearson correlation coefficient as well as the Spearman rank-correlation coefficient to test for monotonicity. Given the exclusively right-handed participants, a net rightward bias ($\beta_R > 0$) was also tested for across each performance group with a one-tailed one-sample *t* test.

For the preferred 2CE1 model and the other 2-parameter GRL models nested within it (2, 2N1, 2N2, 2E1, 2C, 2CN1, and 2CN2), posterior predictive checks were conducted with simulated data sets that were yoked to the empirical data sets and analyzed in the same fashion after averaging across 1,000 simulations. For the first set of checks focusing on only pure GRL ("2") and the full 2CE1 model, participants were initially divided according to the three levels of learning performance. Using one-tailed one-sample *t* tests, above-chance probabilities were tested for with respect to correct actions, right-hand actions, and alternated actions. (By design, alternation of actions was more frequent when actions were more correct.) The net right-hand effects in the Poor-learner and Nonlearner groups were compared to those in the Good-learner group with one-tailed independent-samples *t* tests. Analogous comparisons within and between groups were conducted for the raw measures of absolute lateral bias $|P(Right)\text{-}50\%|$ and absolute repetition-or-alternation frequency $|P(Repeat)\text{-}50\%|$. Moreover, correlations were tested for across the continuous measure of accuracy rather than discrete participant groups.

Individuals across the two learner groups were first reclassified according to the 2CE1 model's fitted result of either leftward bias ($\beta_R < 0$) or rightward bias ($\beta_R > 0$). Above-chance probabilities of either left-hand or right-hand actions were then tested for in empirical data as well as simulated data from the eight 2-parameter GRL models. These individuals were next reclassified according to the 2CE1 model's fitted result of either alternation bias ($\beta_1 < 0$) or repetition bias ($\beta_1 > 0$). (Supplementary analyses further divided six intersectional subgroups as well, crossing the three levels of learning performance with either leftward versus rightward or alternation versus repetition.) The alternation-bias and repetition-bias groups were tested for above-chance probabilities of alternated and repeated actions, respectively. Post-hoc tests followed to check between groups in the event of trending but nonsignificant results within a group—in this case using one-tailed independent-samples *t* tests. The probability of repeating versus alternating was also conditioned on previous actions up to eight trials back. Posterior predictive checks for these action-history curves were generated for both the primary model comparison and the extended model comparison.

For psychometric functions, the first logistic-regression model represented the probability of a right-hand action as a function of the difference between the state-dependent action values $Q_t(s_t,a_R)$ and $Q_t(s_t,a_L)$ corresponding to right and left. The second model represented the probability of repeating the most recent action (independent of state) as a function of the difference between action values that correspond to repetition and alternation. To accommodate

interindividual variability in the range of estimated values, differences in action values were normalized with respect to the maximum absolute value for each participant. Parameters for these mixed-effects models were first estimated at the level of individual participants and then assessed within each bias group by way of one-tailed one-sample *t* tests.

Supporting information

S1 Text. Fig A. Task. This schematic of the hierarchical reversal-learning task performed during fMRI scanning includes the probabilities of a rewarded outcome in one of 12 blocks. Following an intertrial interval (ITI) with a fixation cross, one of four paired states (i.e., cues) was presented with equal probability, prompting the participant to choose either the left-hand action ("L") or the right-hand action ("R"). Confirmation of the action at the reaction time (RT) was followed by an interstimulus interval (ISI) and finally an outcome of either a monetary reward or no reward as feedback. The paired state categories were faces and houses for the 3-T version or colors and directions of motion for the 7-T version. Dotted arrows symbolize the two possible actions. Solid arrows represent equally or more likely state transitions, whereas dashed arrows represent less likely transitions. Arrow thickness corresponds to the weight of an outcome's probability. **(b)** Only one action was rewarded per state, thereby facilitating discriminative action generalization. States were paired within a category as "state A" and "state B" such that opposite actions were rewarded between the two states, thereby facilitating discriminative state generalization. One of two possible arrangements for hierarchical reward structure (independent of probabilities) is shown here, corresponding to the face category for this example block: The upper face is "state A", and the lower face is "state B". There was no pairing between the independent categories. **(c)** The second possible arrangement is also shown for comparison. The two possibilities alternated within categories as this anticorrelational rule remained constant through reversals that remapped categories between blocks. For an optimal learner, this binary metastate determines the cognitive map or model of generalizable task structure, which for a proper (cognitive) model-based algorithm is an explicit model but for generalized reinforcement learning is an implicit model. This figure corresponds to Fig 1 of the original report [12]: https://doi.org/10.1002/hbm.25988.

Fig B in S1 Text. The "generalized reinforcement learning" (GRL) model. Representative dynamics of value signals and learning signals generated by the GRL model are shown for the final participant in the Good-learner group of the 3-T Face/House data set. Parameters were assigned as follows for this participant: $\alpha = 0.318$, $g_A = -0.710$, $g_S = -0.808$, $\tau = 0.408$, $\beta_R = 0.178$, $\beta_I = -0.067$, and $\lambda_H = 0.753$. Tracking the probability of reward for the left and right actions (blue and red lines, respectively) in each of four active states, the model's estimates of action values $Q_t(s,a)$ (solid lines) are plotted along with actual values (dashed lines) over the course of 12 blocks. Plotted below these value signals are time courses of the corresponding action-value-prediction error (AVPE) δ^Q_{t+1} signals, which represent a distinct type of reward-prediction error (RPE) along with the state-value-prediction error (SVPE) δ^V_{t+1} (cf. [12,21]. However, throughout this report, the usage of the generic term "RPE" and its variable "δ_{t+1}" with no superscript—rather than "AVPE" and "δ^Q_{t+1}"—is due to omission of the neural model's SVPE here. Discriminative state and action generalization are evident with counterfactual updates of values for the three nonexperienced state-action pairs within a category. These additional updates occur despite only one state-action pair being experienced with feedback. Each colored tick mark denotes an occurrence of the respective action. This figure corresponds to Fig 7A of the original report [12].

Table A in S1 Text. Model parameters (unrolled). See Table 2. Models are listed individually here.

Fig C in S1 Text. Discriminability of the 2CE1 model: 3-T Face/House version. Compare to Fig 2. Each fitted instantiation of the preferred 2CE1 model was used to simulate a data set yoked to that of the respective participant. The results from the empirical model comparison were replicated in silico as a demonstration of the discriminability of this 7-parameter model among both simpler and more complex alternatives ranging from 0 to 9 free parameters. Model recovery succeeded inasmuch as the 2CE1 model remained preferred among Good learners, and 2CE1 or its nonlearning analog XCE1 could be recovered for Poor learners or Nonlearners as well. See also Tables G, H, and I.

Fig D in S1 Text. Discriminability of the 2CE1 model: 7-T Color/Motion version. Compare to Fig 3 and Fig C. See also Tables J and K.

Fig E in S1 Text. Discriminability of the no-bias model "2" with only GRL: 3-T Face/House version. Compare to Fig C. The no-bias model "2" was recovered in lieu of the bias-and-hysteresis model 2CE1 when substituting data simulated with the no-bias model. This converse model recovery again demonstrates an absence of overfitting. See also Tables L, M, and N.

Fig F in S1 Text. Discriminability of the no-bias model "2" with only GRL: 7-T Color/Motion version. Compare to Figs D and E. See also Tables O and P.

Fig G in S1 Text. Reduced model comparison for discriminability of the 2CE1 model: 3-T Face/House version. Compare to Fig 4 and Fig C.

Fig H in S1 Text. Reduced model comparison for discriminability of the 2CE1 model: 7-T Color/Motion version. Compare to Fig 5 and Figs D and G.

Fig I in S1 Text. Reduced model comparison for discriminability of the no-bias model "2" with only GRL: 3-T Face/House version. Compare to Fig E.

Fig J in S1 Text. Reduced model comparison for discriminability of the no-bias model "2" with only GRL: 7-T Color/Motion version. Compare to Figs F and I.

Fig K in S1 Text. Model comparison by bias category: 3-T Face/House version. Compare to Figs 2 (panel d here) and 4 (a) and Figs C (e), E (f), G (b), and I (c). Participant counts for best-fitting models can also be grouped according to four categories: no bias (e.g., "2"), constant bias (e.g., 2C), hysteretic bias (e.g., 2E1), or both constant and hysteretic bias (e.g., 2CE1).

Fig L in S1 Text. Model comparison by bias category: 7-T Color/Motion version. Compare to Figs 3 (panel d here) and 5 (a) and Figs D (e), F (f), H (b), J (c), and K.

Fig M in S1 Text. Confusion matrix and inverse-confusion matrix. Compare to Figs K and L. The confusion matrix *P(Fit | Simulation)* corresponds to the probability (as a percentage) that simulated data from a given model are best fitted by either the model that actually generated the data or an alternative model. The inversion matrix *P(Simulation | Fit)* instead corresponds to the probability that a model generated the simulated data given that either the same model or an alternative model fitted the data best. **(a)** Limiting the model comparison to only the 2CE1 or "2" models with or without bias and hysteresis, model confusion is minimal as expected. **(b-c)** Expanding the model comparison with a binarized categorization of bias versus none for either 8 (b) or 72 (c) models does leave confusion less minimal as the models with bias outnumber the models without bias, but the expected trend of model recovery still holds true. **(d-f)** Results were replicated in the 7-T Color/Motion version of the experiment.

Fig N in S1 Text. Parameter recovery with the 2CE1 model more accurate than recovery with the no-bias model "2" including only GRL. (a) As described previously, the 2CE1 model was fitted to yoked simulated data that were generated with the 2CE1 parameters originally fitted to empirical data. Parameter recovery was especially robust for the Good-learner group across all seven free parameters, including β_R, β_I, and λ_H for action bias and hysteresis ($p < 0.05$). Although somewhat less robust, recovery of 2CE1 parameters was also successful for the Poor-learner group ($p < 0.05$ with the exception of τ from the first data set and $p < 0.06$ for α from the second data set). **(b)** The relative significance of bias and hysteresis was found to be greatest

among Poor learners. Hence, if instead fitting the no-bias model "2", the remaining four parameters needed for pure GRL (α, g_A, g_S, τ) were not significantly recoverable for the Poor-learner group ($p > 0.05$ with one exception for g_S from the second data set). **(c)** The p values for the correlations are plotted separately for Good ("G") and Poor ("P") learners when using either the 7-parameter 2CE1 model or the 4-parameter model "2". **(d-f)** Results were replicated in the 7-T Color/Motion version of the experiment.

Fig O in S1 Text. Action bias and hysteresis versus learning performance: Individual results. Compare to Figs 6 and 7.

Fig P in S1 Text. Hysteresis represented by sequences across trials. Compare to Fig 12. The distribution of lengths of runs of consecutive repeated actions reveals hysteresis from another perspective. Alternation and repetition biases should result in shorter and longer runs, respectively, as only a model including hysteresis could replicate. Error bars indicate standard errors of the means.

Fig Q in S1 Text. Constant bias and learning performance. Compare to Figs 6, 7, and 8. Participants were further divided into six subgroups that separated the two directions of constant lateral bias as well as the three levels of learning performance. Constant bias should still be substantial for Good learners but should be even more pronounced for Poor learners and Nonlearners. Moreover, modeled bias in 2CE1 simulations should still both qualitatively and quantitively replicate the directions and magnitudes of empirical effects of laterality.

Fig R in S1 Text. Hysteresis and learning performance. Compare to Figs 6, 7, and 9. Participants were next subdivided with the two directions of hysteretic bias as the first factor crossed with learning performance. As with constant bias, hysteretic bias should still be substantial for Good learners but should be even more pronounced for Poor learners and Nonlearners. Likewise, modeled bias in 2CE1 simulations should still replicate the directions and magnitudes of empirical effects of hysteresis.

Fig S in S1 Text. Substitution of state-dependent action hysteresis. Compare to Figs 12 and 14. The alternative of state-dependent hysteresis $H_t(s_t,a)$ was first substituted in place of state-independent hysteresis $H_t(a)$. Following the original reduced comparison of eight models, here state-dependent action hysteresis was tested in its 1-back (2CsN1), 2-back (2CsN2) and exponential (2CsE1) forms. As expected because the four states (which this hysteresis depends on) were rotated in sequence, each form of state-dependent hysteresis by itself failed to match the action-history curves here.

Fig T in S1 Text. Addition of state-dependent action hysteresis to state-independent action hysteresis. Compare to Figs 12 and 14 and Fig S. State-dependent hysteresis $H_t(s_t,a)$ in exponential form ("sE1+") was subsequently added to the eight models from the original reduced model comparison with state-independent hysteresis $H_t(a)$ (2 through 2CE1). Considering that the 2CE1 model in its own right could parsimoniously account for all of these alternation and repetition effects, the expanded sE1+2CE1 model was not justified by any qualitative improvement in fit.

Fig U in S1 Text. Addition of state-independent action value. Compare to Figs 12 and 14. State-independent action value $Q_t(a)$ was added to the eight models from the original reduced model comparison with only state-dependent action value $Q_t(s_t,a)$. Again, the expanded Qa+2CE1 model was not justified by any qualitative improvement in fit.

Fig V in S1 Text. Addition of confirmation bias. Compare to Figs 12 and 14. A second learning rate was added to distinguish updates for positive and negative reward-prediction errors (α_P and α_N). Models with confirmation bias ("cLR+") in particular imposed the constraint $\alpha_N < \alpha_P$ with an assumption of subjective optimism biased toward positive valence. The expanded cLR+2CE1 model was not justified by any qualitative improvement in fit.

Fig W in S1 Text. Addition of asymmetric learning rates. Compare to Figs 12 and 14 and Fig V. As before, a second learning rate was added to distinguish updates for positive and negative

reward-prediction errors, but here the asymmetric learning rates $\alpha_N \neq \alpha_P$ had no constraint of confirmation bias such that pessimistic doubt bias was also a possibility. Even in this unconstrained form, the expanded LR+2CE1 model still was not justified by any qualitative improvement in fit.

Table B in S1 Text. Model comparison: 3-T Face/House version (Good-learner group). See Fig 2. Listed first for the 72 models fitted to empirical data are absolute scores for deviance and the corrected Akaike information criterion (AICc), where a lower score is better. These absolute scores were translated to residual goodness of fit relative to the null chance model "X", where a higher score is better. Results with the best fit according to the AICc, which penalizes degrees of freedom, are highlighted with boldface and italics. "df" stands for degrees of freedom. The conventions for displaying this table also apply for Tables C-U.

Table C in S1 Text. Model comparison: 3-T Face/House version (Poor-learner group). See Fig 2.

Table D in S1 Text. Model comparison: 3-T Face/House version (Nonlearner group). See Fig 2.

Table E in S1 Text. Model comparison: 7-T Color/Motion version (Good-learner group). See Fig 3.

Table F in S1 Text. Model comparison: 7-T Color/Motion version (Poor-learner group). See Fig 3.

Table G in S1 Text. Discriminability of the 2CE1 model: 3-T Face/House version (Good-learner group). See Fig C.

Table H in S1 Text. Discriminability of the 2CE1 model: 3-T Face/House version (Poor-learner group). See Fig C.

Table I in S1 Text. Discriminability of the 2CE1 model: 3-T Face/House version (Nonlearner group). See Fig C.

Table J in S1 Text. Discriminability of the 2CE1 model: 7-T Color/Motion version (Good-learner group). See Fig D.

Table K in S1 Text. Discriminability of the 2CE1 model: 7-T Color/Motion version (Poor-learner group). See Fig D.

Table L in S1 Text. Discriminability of the no-bias model "2" with only GRL: 3-T Face/House version (Good-learner group). See Fig E.

Table M in S1 Text. Discriminability of the no-bias model "2" with only GRL: 3-T Face/House version (Poor-learner group). See Fig E.

Table N in S1 Text. Discriminability of the no-bias model "2" with only GRL: 3-T Face/House version (Nonlearner group). See Fig E.

Table O in S1 Text. Discriminability of the no-bias model "2" with only GRL: 7-T Color/Motion version (Good-learner group). See Fig F.

Table P in S1 Text. Discriminability of the no-bias model "2" with only GRL: 7-T Color/Motion version (Poor-learner group). See Fig F.

Table Q in S1 Text. Extended model comparison: 3-T Face/House version (Good-learner group). See Table 4. Results with the best fit within each subset of 8 models are highlighted with boldface and italics. Results with the best fit across all 44 models are also marked with asterisks.

Table R in S1 Text. Extended model comparison: 3-T Face/House version (Poor-learner group). See Table 4.

Table S in S1 Text. Extended model comparison: 3-T Face/House version (Nonlearner group). See Table 4.

Table T in S1 Text. Extended model comparison: 7-T Color/Motion version (Good-learner group). See Table 4.

Table U in S1 Text. Extended model comparison: 7-T Color/Motion version (Poor-learner group). See Table 4.
(PDF)

S1 Data. Data. All data are included.
(ZIP)

Acknowledgments

We thank the other coauthors of the previous report of this study: Neil Dundon, Raphael Gerraty, Natalie Saragosa-Harris, Karol Szymula, Koranis Tanwisuth, Michael Tyszka, Camilla van Geen, Harang Ju, Arthur Toga, Joshua Gold, Dani Bassett, Catherine Hartley, and Daphna Shohamy.

Author Contributions

Conceptualization: Jaron T. Colas.

Formal analysis: Jaron T. Colas.

Funding acquisition: John P. O'Doherty, Scott T. Grafton.

Investigation: Jaron T. Colas.

Methodology: Jaron T. Colas.

Supervision: John P. O'Doherty, Scott T. Grafton.

Visualization: Jaron T. Colas.

Writing – original draft: Jaron T. Colas, John P. O'Doherty, Scott T. Grafton.

References

1. Bush RR, Mosteller F. A mathematical model for simple learning. Psychol Rev. 1951; 58(5): 313–323. https://doi.org/10.1037/h0054388 PMID: 14883244
2. Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, editors. Classical conditioning II: Current research and theory. New York (NY): Appleton-Century-Crofts; 1972. p. 64–99.
3. Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge (MA): MIT Press; 1998.
4. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci. 2017; 12(6): 1100–1122. https://doi.org/10.1177/1745691617693393 PMID: 28841086
5. Plonsky O, Apel R, Ert E, Tennenholtz M, Bourgin D, Peterson JC, Reichman D, Griffiths TL, Russell SJ, Carter EC, Cavanagh JF, Erev I. Predicting human decisions with behavioral theories and machine learning. arXiv. 2019; 1904.06866. https://doi.org/10.48550/arxiv.1904.06866
6. O'Doherty JP, Cockburn J, Pauli WM. Learning, reward, and decision making. Annu Rev Psychol. 2017; 68(1): 73–100. https://doi.org/10.1146/annurev-psych-010416-044216 PMID: 27687119
7. O'Doherty JP, Lee S, Tadayonnejad R, Cockburn J, Iigaya K, Charpentier CJ. Why and how the brain weights contributions from a mixture of experts. Neurosci Biobehav Rev. 2021; 123: 14–23. https://doi.org/10.1016/j.neubiorev.2020.10.022 PMID: 33444700
8. Averbeck B, O'Doherty JP. Reinforcement-learning in fronto-striatal circuits. Neuropsychopharmacology. 2022; 47(1): 147–162. https://doi.org/10.1038/s41386-021-01108-0 PMID: 34354249
9. Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010; 66(4): 585–595. https://doi.org/10.1016/j.neuron.2010.04.016 PMID: 20510862
10. Momennejad I, Russek EM, Cheong JH, Botvinick MM, Daw ND, Gershman SJ. The successor representation in human reinforcement learning. Nat Hum Behav. 2017; 1: 680–692. https://doi.org/10.1038/s41562-017-0180-8 PMID: 31024137

11. Eckstein MK, Collins AG. Computational evidence for hierarchically structured reinforcement learning in humans. Proc Natl Acad Sci U S A. 2020; 117(47): 29381–29389. https://doi.org/10.1073/pnas.1912330117 PMID: 33229518
12. Colas JT, Dundon NM, Gerraty RT, Saragosa-Harris NM, Szymula KP, Tanwisuth K, Tyszka JM, van Geen C, Ju H, Toga AW, Gold JI, Bassett DS, Hartley CA, Shohamy D, Grafton ST, O'Doherty JP. Reinforcement learning with associative or discriminative generalization across states and actions: fMRI at 3 T and 7 T. Hum Brain Mapp. 2022; 43(15): 4750–4790. https://doi.org/10.1002/hbm.25988
13. Daw ND. Trial-by-trial data analysis using computational models. In: Delgado MR, Phelps EA, Robbins TW, editors. Decision making, affect, and learning: attention and performance XXIII. New York (NY): Oxford University Press; 2011. p. 3–38. https://doi.org/10.1093/acprof:oso/9780199600434.001.0001
14. Hampton AN, Bossaerts P, O'Doherty JP. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci. 2006; 26(32): 8360–8367. https://doi.org/10.1523/JNEUROSCI.1010-06.2006 PMID: 16899731
15. Hampton AN, Adolphs R, Tyszka JM, O'Doherty JP. Contributions of the amygdala to reward expectancy and choice signals in human prefrontal cortex. Neuron. 2007; 55(4): 545–555. https://doi.org/10.1016/j.neuron.2007.07.022 PMID: 17698008
16. Gläscher J, Hampton AN, O'Doherty JP. Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. Cereb Cortex. 2009; 19 (2): 483–495. https://doi.org/10.1093/cercor/bhn098 PMID: 18550593
17. Lau B, Glimcher PW. Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav. 2005; 84(3): 555–579. https://doi.org/10.1901/jeab.2005.110-04 PMID: 16596980
18. Schönberg T, Daw ND, Joel D, O'Doherty JP. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci. 2007; 27 (47): 12860–12867. https://doi.org/10.1523/JNEUROSCI.2496-07.2007 PMID: 18032658
19. Gershman SJ, Pesaran B, Daw ND. Human reinforcement learning subdivides structured action spaces by learning effector-specific values. J Neurosci. 2009; 29(43): 13524–13531. https://doi.org/10.1523/JNEUROSCI.2469-09.2009 PMID: 19864565
20. Rutledge RB, Lazzaro SC, Lau B, Myers CE, Gluck MA, Glimcher PW. Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. J Neurosci. 2009; 29(48): 15104–15114. https://doi.org/10.1523/JNEUROSCI.3524-09.2009 PMID: 19955362
21. Colas JT, Pauli WM, Larsen T, Tyszka JM, O'Doherty JP. Distinct prediction errors in mesostriatal circuits of the human brain mediate learning about the values of both states and actions: evidence from high-resolution fMRI. PLOS Comput Biol. 2017; 13(10): e1005810. https://doi.org/10.1371/journal.pcbi.1005810 PMID: 29049406
22. O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science. 2004; 304(5669): 452–454. https://doi.org/10.1126/science.1094285 PMID: 15087550
23. Shepard RN. Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. Psychometrika. 1957; 22(4): 325–345. https://doi.org/10.1007/bf02288967
24. Luce RD. Individual choice behavior: a theoretical analysis. New York (NY): Wiley; 1959. https://doi.org/10.1037/14396-000
25. Luce RD. The choice axiom after twenty years. J Math Psychol. 1977; 15(3): 215–233. https://doi.org/10.1016/0022-2496(77)90032-3
26. Busemeyer JR, Diederich A. Cognitive modeling. Thousand Oaks (CA): Sage; 2010.
27. Palminteri S, Wyart V, Koechlin E. The importance of falsification in computational cognitive modeling. Trends Cogn Sci. 2017; 21(6): 425–433. https://doi.org/10.1016/j.tics.2017.03.011 PMID: 28476348
28. Wilson RC, Collins AG. Ten simple rules for the computational modeling of behavioral data. eLife. 2019; 8: e49547. https://doi.org/10.7554/eLife.49547 PMID: 31769410
29. Wiecki TV, Poland J, Frank MJ. Model-based cognitive neuroscience approaches to computational psychiatry: clustering and classification. Clinical Psychol Sci. 2015; 3(3): 378–399. https://doi.org/10.1177/2167702614565359
30. Scheibehenne B, Pachur T. Hierarchical Bayesian modeling: does it improve parameter stability? In: Knauff M, Pauen M, Sebanz N, Wachsmuth I, editors. Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin (TX): Cognitive Science Society; 2013. p. 1277–1282.
31. Scheibehenne B, Pachur T. Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. Psychon Bull Rev. 2015; 22: 391–407. https://doi.org/10.3758/s13423-014-0684-4 PMID: 25134469

32. Spektor MS, Kellen D. The relative merit of empirical priors in non-identifiable and sloppy models: applications to models of learning and decision-making. Psychon Bull Rev. 2018; 25(6): 2047–2068. https://doi.org/10.3758/s13423-018-1446-5

33. Ballard IC, McClure SM. Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. J Neurosci Methods. 2019; 317: 37–44. https://doi.org/10.1016/j.jneumeth.2019.01.006 PMID: 30664916

34. Katahira K, Oba T, Toyama A. Can reliability of computational models be truly improved by placing priors on parameters? PsyArXiv. 2022; 4e2d9. https://doi.org/10.31234/osf.io/4e2d9

35. Baribault B, Collins AG. Troubleshooting Bayesian cognitive models. Psychol Methods. 2023. https://doi.org/10.1037/met0000554

36. Thorndike EL. On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. Am J Psychol. 1939; 52(1): 122–124. https://doi.org/10.2307/1416673

37. Robinson WS. Ecological correlations and the behavior of individuals. Am Sociol Rev. 1950; 15(3): 351–357. https://doi.org/10.2307/2087176

38. Selvin HC. Durkheim's Suicide and problems of empirical research. Am J Sociol. 1958; 63(6): 607–619. https://doi.org/10.1086/222356

39. Corrado GS, Sugrue LP, Seung HS, Newsome WT. Linear-nonlinear-Poisson models of primate choice dynamics. J Exp Anal Behav. 2005; 84(3): 581–617. https://doi.org/10.1901/jeab.2005.23-05 PMID: 16596981

40. Katahira K. The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. J Math Psychol. 2015; 66: 59–69. https://doi.org/10.1016/j.jmp.2015.03.006

41. Katahira K, Bai Y, Nakao T. Pseudo-learning effects in reinforcement learning model-based analysis: a problem of misspecification of initial preference. PsyArXiv. 2017; a6hzq. https://doi.org/10.31234/osf.io/a6hzq

42. Katahira K. The statistical structures of reinforcement learning with asymmetric value updates. J Math Psychol. 2018; 87: 31–45. https://doi.org/10.1016/j.jmp.2018.09.002

43. Toyama A, Katahira K, Ohira H. Biases in estimating the balance between model-free and model-based learning systems due to model misspecification. J Math Psychol. 2019; 91: 88–102. https://doi.org/10.1016/j.jmp.2019.03.007

44. Sugawara M, Katahira K. Dissociation between asymmetric value updating and perseverance in human reinforcement learning. Sci Rep. 2021; 11: 3574. https://doi.org/10.1038/s41598-020-80593-7 PMID: 33574424

45. Katahira K, Kimura K. Influences of reinforcement and choice histories on choice behavior in actor-critic learning. Comput Brain Behav. 2023; 6: 172–194. https://doi.org/10.1007/s42113-022-00145-2

46. Palminteri S. Choice-confirmation bias and gradual perseveration in human reinforcement learning. Behav Neurosci. 2023; 137(1): 78–88. https://doi.org/10.1037/bne0000541 PMID: 36395020

47. Toyama A, Katahira K, Kunisato Y. Examinations of biases by model misspecification and parameter reliability of reinforcement learning models. Comput Brain Behav. 2023; 6: 651–670. https://doi.org/10.1007/s42113-023-00175-4

48. Myung IJ. The importance of complexity in model selection. J Math Psychol. 2000; 44(1): 190–204. https://doi.org/10.1006/jmps.1999.1283 PMID: 10733864

49. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Comput. 1991; 3(1): 79–87. https://doi.org/10.1162/neco.1991.3.1.79 PMID: 31141872

50. Doya K, Samejima K, Katagiri KI, Kawato M. Multiple model-based reinforcement learning. Neural Comput. 2002; 14(6): 1347–1369. https://doi.org/10.1162/089976602753712972 PMID: 12020450

51. Yuksel SE, Wilson JN, Gader PD. Twenty years of mixture of experts. IEEE Trans Neural Netw Learn Syst. 2012; 23(8): 1177–1193. https://doi.org/10.1109/TNNLS.2012.2200299 PMID: 24807516

52. Hamrick JB, Ballard AJ, Pascanu R, Vinyals O, Heess N, Battaglia PW. Metacontrol for adaptive imagination-based optimization. arXiv. 2017; 1705.02670. https://doi.org/10.48550/arxiv.1705.02670

53. Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, Dean J. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. arXiv. 2017; 1701.06538. https://doi.org/10.48550/arxiv.1701.06538

54. Fedus W, Dean J, Zoph B. A review of sparse expert models in deep learning. arXiv. 2022; 2209.01667. https://doi.org/10.48550/arxiv.2209.01667

55. Graybiel AM, Aosaki T, Flaherty AW, Kimura M. The basal ganglia and adaptive motor control. Science. 1994; 265(5180): 1826–1831. https://doi.org/10.1126/science.8091209 PMID: 8091209

56. Ghahramani Z, Wolpert DM. Modular decomposition in visuomotor learning. Nature. 1997; 386 (6623): 392–395. https://doi.org/10.1038/386392a0 PMID: 9121554

57. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005; 8(12): 1704–1711. https://doi.org/10.1038/nn1560 PMID: 16286932

58. Ito M, Doya K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J Neurosci. 2009; 29(31): 9861–9874. https://doi.org/10.1523/JNEUROSCI.6157-08.2009 PMID: 19657038

59. Kim H, Sul JH, Huh N, Lee D, Jung MW. Role of striatum in updating values of chosen actions. J Neurosci. 2009; 29(47): 14701–14712. https://doi.org/10.1523/JNEUROSCI.2728-09.2009 PMID: 19940165

60. Fonseca MS, Murakami M, Mainen ZF. Activation of dorsal raphe serotonergic neurons promotes waiting but is not reinforcing. Curr Biol. 2015; 25(3): 306–315. https://doi.org/10.1016/j.cub.2014.12.002 PMID: 25601545

61. Beron CC, Neufeld SQ, Linderman SW, Sabatini BL. Mice exhibit stochastic and efficient action switching during probabilistic decision making. Proc Natl Acad Sci U S A. 2022; 119(15): e2113961119. https://doi.org/10.1073/pnas.2113961119 PMID: 35385355

62. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr. 1974; 19(6): 716–723. https://doi.org/10.1109/tac.1974.1100705

63. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika. 1989; 76(2): 297–307. https://doi.org/10.1093/biomet/76.2.297

64. Thorndike EL. The fundamentals of learning. New York (NY): Teachers College Bureau of Publications, Columbia University; 1932. https://doi.org/10.1037/10976-000

65. Thorndike EL. A proof of the law of effect. Science. 1933; 77(1989): 173–175. https://doi.org/10.1126/science.77.1989.173-a

66. Frank MJ, Seeberger LC, O'Reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004; 306(5703): 1940–1943. https://doi.org/10.1126/science.1102941 PMID: 15528409

67. Sharot T. The optimism bias. Curr Biol. 2011; 21(23): R941–R945. https://doi.org/10.1016/j.cub.2011.10.030 PMID: 22153158

68. Sharot T, Korn CW, Dolan RJ. How unrealistic optimism is maintained in the face of reality. Nat Neurosci. 2011; 14(11): 1475–1479. https://doi.org/10.1038/nn.2949 PMID: 21983684

69. Daw ND, Kakade S, Dayan P. Opponent interactions between serotonin and dopamine. Neural Netw. 2002; 15(4–6): 603–616. https://doi.org/10.1016/s0893-6080(02)00052-7 PMID: 12371515

70. Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A. 2007; 104(41): 16311–16316. https://doi.org/10.1073/pnas.0706111104 PMID: 17913879

71. Frank MJ, Doll BB, Oas-Terpstra J, Moreno F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci. 2009; 12(8): 1062–1068. https://doi.org/10.1038/nn.2342 PMID: 19620978

72. Niv Y, Edlund JA, Dayan P, O'Doherty JP. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. J Neurosci. 2012; 32(2): 551–562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012 PMID: 22238090

73. Cazé RD, van der Meer MA. Adaptive properties of differential learning rates for positive and negative outcomes. Biol Cybern. 2013; 107(6): 711–719. https://doi.org/10.1007/s00422-013-0571-5 PMID: 24085507

74. Lefebvre G, Lebreton M, Meyniel F, Bourgeois-Gironde S, Palminteri S. Behavioural and neural characterization of optimistic reinforcement learning. Nat Hum Behav. 2017; 1(4): 0067. https://doi.org/10.1038/s41562-017-0067

75. Palminteri S, Lefebvre G, Kilford EJ, Blakemore SJ. Confirmation bias in human reinforcement learning: evidence from counterfactual feedback processing. PLOS Comput Biol. 2017; 13(8): e1005684. https://doi.org/10.1371/journal.pcbi.1005684 PMID: 28800597

76. Palminteri S, Lebreton M. The computational roots of positivity and confirmation biases in reinforcement learning. Trends Cogn Sci. 2022; 26(7): 607–621. https://doi.org/10.1016/j.tics.2022.04.005 PMID: 35662490

77. Salem-Garcia N, Palminteri S, Lebreton M. Linking confidence biases to reinforcement-learning processes. Psychol Rev. 2023; 130(4): 1017–1043. https://doi.org/10.1037/rev0000424 PMID: 37155268

78. Ting CC, Salem-Garcia N, Palminteri S, Engelmann JB, Lebreton M. Neural and computational underpinnings of biased confidence in human reinforcement learning. Nat Commun. 2023; 14(1): 6896. https://doi.org/10.1038/s41467-023-42589-5 PMID: 37898640

79. Gershman SJ. Empirical priors for reinforcement learning models. J Math Psychol. 2016; 71: 1–6. https://doi.org/10.1016/j.jmp.2016.01.006

80. Chambon V, Théro H, Vidal M, Vandendriessche H, Haggard P, Palminteri S. Information about action outcomes differentially affects learning from self-determined versus imposed choices. Nat Hum Behav. 2020; 4(10): 1067–1079. https://doi.org/10.1038/s41562-020-0919-5 PMID: 32747804

81. Findling C, Hubert F, International Brain Laboratory, Acerbi L, Benson B, Benson J, Birman D, Bonacchi N, Carandini M, Catarino JA, Chapuis GA, Churchland AK, Dan Y, DeWitt EE, Engel TA, Fabbri M, Faulkner M, Fiete IR, Freitas-Silva L, Gerçek B, Harris KD, Häusser M, Hofer SB, Hu F, Huntenburg JM, Khanal A, Krasniak C, Langdon C, Latham PE, Lau PY, Mainen Z, Meijer GT, Miska NJ, Mrsic-Flogel TD, Noel J, Nylund K, Pan-Vazquez A, Paniniski L, Pillow J, Rossant C, Roth N, Schaeffer R, Schartner M, Shi Y, Socha KZ, Steinmetz NA, Svoboda K, Tessereau C, Urai AE, Wells MJ, West SJ, Whiteway MR, Winter O, Witten IB, Zador A, Dayan P, Pouget A. Brain-wide representations of prior information in mouse decision-making. bioRxiv. 2023; 547684. https://doi.org/10.1101/2023.07.04.547684

82. Behrens TE, Woolrich MW, Walton ME, Rushworth MF. Learning the value of information in an uncertain world. Nat Neurosci. 2007; 10(9): 1214–1221. https://doi.org/10.1038/nn1954 PMID: 17676057

83. Krugel LK, Biele G, Mohr PN, Li SC, Heekeren HR. Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. Proc Natl Acad Sci U S A. 2009; 106 (42): 17951–17956. https://doi.org/10.1073/pnas.0905191106 PMID: 19822738

84. Nassar MR, Wilson RC, Heasly B, Gold JI. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J Neurosci. 2010; 30(37): 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010 PMID: 20844132

85. Nassar MR, Gold JI. A healthy fear of the unknown: perspectives on the interpretation of parameter fits from computational models in neuroscience. PLOS Comput Biol. 2013; 9(4): e1003015. https://doi.org/10.1371/journal.pcbi.1003015 PMID: 23592963

86. Bak JH, Choi JY, Akrami A, Witten I, Pillow JW. Adaptive optimal training of animal behavior. Adv Neural Inf Process Syst. 2016; 29.

87. Nassar MR, Frank MJ. Taming the beast: extracting generalizable knowledge from computational models of cognition. Curr Opin Behav Sci. 2016; 11: 49–54. https://doi.org/10.1016/j.cobeha.2016.04.003 PMID: 27574699

88. Roy NA, Bak JH, Akrami A, Brody C, Pillow JW. Efficient inference for time-varying behavior during learning. Adv Neural Inf Process Syst. 2018; 31. PMID: 31244514

89. Roy NA, Bak JH, International Brain Laboratory, Akrami A, Brody CD, Pillow JW. Extracting the dynamics of behavior in sensory decision-making experiments. Neuron. 2021; 109(4): 597–610. https://doi.org/10.1016/j.neuron.2020.12.004

90. Ashwood ZC, Roy NA, Stone IR, International Brain Laboratory, Urai AE, Churchland AK, Pouget A, Pillow JW. Mice alternate between discrete strategies during perceptual decision-making. Nat Neurosci. 2022; 25(2): 201–212. https://doi.org/10.1038/s41593-021-01007-z

91. Maggi S, Hock RM, O'Neill M, Buckley MJ, Moran PM, Bast T, Sami M, Humphries MD. Tracking subjects' strategies in behavioural choice experiments at trial resolution. eLife. 2024; 13: e86491. https://doi.org/10.7554/eLife.86491

92. Bruijns SA, International Brain Laboratory, Bougrova K, Laranjeira IC, Lau PY, Meijer GT, Miska NJ, Noel J, Pan-Vazquez A, Roth N, Socha KZ, Urai AE, Dayan P. Dissecting the complexities of learning with infinite hidden Markov models. bioRxiv. 2023; 573001. https://doi.org/10.1101/2023.12.22.573001

93. Le NM, Yildirim M, Wang Y, Sugihara H, Jazayeri M, Sur M. Mixtures of strategies underlie rodent behavior during reversal learning. PLOS Comput Biol. 2023; 19(9): e1011430. https://doi.org/10.1371/journal.pcbi.1011430 PMID: 37708113

94. Miller KJ, Botvinick MM, Brody CD. Dorsal hippocampus contributes to model-based planning. Nat Neurosci. 2017; 20(9): 1269–1276. https://doi.org/10.1038/nn.4613 PMID: 28758995

95. Shahar N, Moran R, Hauser TU, Kievit RA, McNamee D, Moutoussis M, NSPN Consortium, Dolan RJ. Credit assignment to state-independent task representations and its relationship with model-based decision making. Proc Natl Acad Sci U S A. 2019; 116(32): 15871–15876. https://doi.org/10.1073/pnas.1821647116

96. Miller KJ, Botvinick MM, Brody CD. From predictive models to cognitive models: separable behavioral processes underlying reward learning in the rat. bioRxiv. 2018; 461129. https://doi.org/10.1101/461129

97. Shahar N, Hauser TU, Moran R, Moutoussis M, NSPN Consortium, Bullmore ET, Dolan RJ. Assigning the right credit to the wrong action: compulsivity in the general population is associated with augmented outcome-irrelevant value-based learning. Transl Psychiatry. 2021; 11(1): 1–9. https://doi.org/10.1038/s41398-021-01642-x

98. Miller KJ, Botvinick MM, Brody CD. Value representations in the rodent orbitofrontal cortex drive learning, not choice. eLife. 2022; 11: e64575. https://doi.org/10.7554/eLife.64575 PMID: 35975792

99. Yi S, O'Doherty JP. Computational and neural mechanisms underlying the influence of action affordances on value-based choice. bioRxiv. 2023; 550102. https://doi.org/10.1101/2023.07.21.550102

100. Guitart-Masip M, Huys QJ, Fuentemilla L, Dayan P, Duzel E, Dolan RJ. Go and no-go learning in reward and punishment: interactions between affect and effect. NeuroImage. 2012; 62(1): 154–166. https://doi.org/10.1016/j.neuroimage.2012.04.024 PMID: 22548809

101. Guitart-Masip M, Economides M, Huys QJ, Frank MJ, Chowdhury R, Duzel E, Dayan P, Dolan RJ. Differential, but not opponent, effects of L-DOPA and citalopram on action learning with reward and punishment. Psychopharmacology. 2014; 231(5): 955–966. https://doi.org/10.1007/s00213-013-3313-4 PMID: 24232442

102. Millner AJ, Gershman SJ, Nock MK, den Ouden HE. Pavlovian control of escape and avoidance. J Cogn Neurosci. 2018; 30(10): 1379–1390. https://doi.org/10.1162/jocn_a_01224 PMID: 29244641

103. Gershman SJ, Guitart-Masip M, Cavanagh JF. Neural signatures of arbitration between Pavlovian and instrumental action selection. PLOS Comput Biol. 2021; 17(2): e1008553. https://doi.org/10.1371/journal.pcbi.1008553 PMID: 33566831

104. Weber ID, Zorowitz S, Niv Y, Bennett D. The effects of induced positive and negative affect on Pavlovian-instrumental interactions. Cogn Emot. 2022; 36(7): 1343–1360. https://doi.org/10.1080/02699931.2022.2109600 PMID: 35929878

105. Zorowitz S, Karni G, Paredes N, Daw N, Niv Y. Improving the reliability of the Pavlovian go/no-go task. PsyArXiv. 2023; eb697. https://doi.org/10.31234/osf.io/eb697

106. Colas JT, Lu J. Learning where to look for high value improves decision making asymmetrically. Front Psychol. 2017; 8: 2000. https://doi.org/10.3389/fpsyg.2017.02000 PMID: 29187831

107. Voss A, Voss J, Klauer KC. Separating response-execution bias from decision bias: arguments for an additional parameter in Ratcliff's diffusion model. Br J Math Stat Psychol. 2010; 63(3): 539–555. https://doi.org/10.1348/000711009X477581 PMID: 20030967

108. Busse L, Ayaz A, Dhruv NT, Katzner S, Saleem AB, Schölvinck ML, Zaharia AD, Carandini M. The detection of visual contrast in the behaving mouse. J Neurosci. 2011; 31(31): 11351–11361. https://doi.org/10.1523/JNEUROSCI.6689-10.2011 PMID: 21813694

109. Treviño M. Stimulus similarity determines the prevalence of behavioral laterality in a visual discrimination task for mice. Sci Rep. 2014; 4(1): 1–12. https://doi.org/10.1038/srep07569

110. Treviño M, Medina-Coss y León R. Distributed processing of side-choice biases. Brain Res. 2020; 1749: 147138. https://doi.org/10.1016/j.brainres.2020.147138 PMID: 33002485

111. Treviño M, Medina-Coss y León R, Haro B. Adaptive choice biases in mice and humans. Front Behav Neurosci. 2020; 14: 99. https://doi.org/10.3389/fnbeh.2020.00099 PMID: 32760255

112. Treviño M, Castiello S, Arias-Carrión O, De la Torre-Valdovinos B, Medina-Coss y León R. Isomorphic decisional biases across perceptual tasks. PLOS ONE. 2021; 16(1): e0245890. https://doi.org/10.1371/journal.pone.0245890 PMID: 33481948

113. Dundon NM, Colas JT, Garrett N, Babenko V, Rizor E, Yang D, MacNamara M, Petzold L, Grafton ST. Decision heuristics in contexts integrating action selection and execution. Sci Rep. 2023; 13: 6486. https://doi.org/10.1038/s41598-023-33008-2 PMID: 37081031

114. Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia. 1971; 9(1): 97–113. https://doi.org/10.1016/0028-3932(71)90067-4 PMID: 5146491

115. Amunts K, Schlaug G, Schleicher A, Steinmetz H, Dabringhaus A, Roland PE, Zilles K. Asymmetry in the human motor cortex and handedness. NeuroImage. 1996; 4(3): 216–222. https://doi.org/10.1006/nimg.1996.0073 PMID: 9345512

116. Amunts K, Jäncke L, Mohlberg H, Steinmetz H, Zilles K. Interhemispheric asymmetry of the human motor cortex related to handedness and gender. Neuropsychologia. 2000; 38(3): 304–312. https://doi.org/10.1016/s0028-3932(99)00075-5 PMID: 10678696

117. Schmidt SL, Oliveira RM, Krahe TE, Filgueiras CC. The effects of hand preference and gender on finger tapping performance asymmetry by the use of an infra-red light measurement device. Neuropsychologia. 2000; 38(5): 529–534. https://doi.org/10.1016/s0028-3932(99)00120-7 PMID: 10689030

118. Grafton ST, Hazeltine E, Ivry RB. Motor sequence learning with the nondominant left hand. Exp Brain Res. 2002; 146(3): 369–378. https://doi.org/10.1007/s00221-002-1181-y

119. Krajbich I, Armel C, Rangel A. Visual fixations and the computation and comparison of value in simple choice. Nat Neurosci. 2010; 13(10): 1292–1298. https://doi.org/10.1038/nn.2635

120. Krajbich I, Rangel A. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. Proc Natl Acad Sci U S A. 2011; 108(33): 13852–13857. https://doi.org/10.1073/pnas.1101328108 PMID: 21808009

121. Reutskaja E, Nagel R, Camerer CF, Rangel A. Search dynamics in consumer choice under time pressure: an eye-tracking study. Am Econ Rev. 2011; 101(2): 900–926. https://doi.org/10.1257/aer.101.2.900

122. Jewell G, McCourt ME. Pseudoneglect: a review and meta-analysis of performance factors in line bisection tasks. Neuropsychologia. 2000; 38(1): 93–110. https://doi.org/10.1016/s0028-3932(99)00045-7 PMID: 10617294

123. Heilman KM, Van Den Abell T. Right hemisphere dominance for attention: the mechanism underlying hemispheric asymmetries of inattention (neglect). Neurology. 1980; 30(3): 327–330. https://doi.org/10.1212/wnl.30.3.327 PMID: 7189037

124. Mesulam MM. A cortical network for directed attention and unilateral neglect. Ann Neurol. 1981; 10(4): 309–325. https://doi.org/10.1002/ana.410100402 PMID: 7032417

125. Vallortigara G. The evolutionary psychology of left and right: costs and benefits of lateralization. Dev Psychobiol. 2006; 48(6): 418–427. https://doi.org/10.1002/dev.20166 PMID: 16886183

126. de Schotten MT, Dell'Acqua F, Forkel SJ, Simmons A, Vergani F, Murphy DG, Catani M. A lateralized brain network for visuospatial attention. Nat Neurosci. 2011; 14(10): 1245–1246. https://doi.org/10.1038/nn.2905 PMID: 21926985

127. Chokron S, Imbert M. Influence of reading habits on line bisection. Cogn Brain Res. 1993; 1(4): 219–222. https://doi.org/10.1016/0926-6410(93)90005-p PMID: 8003920

128. Chokron S, De Agostini M. Reading habits and line bisection: a developmental approach. Cogn Brain Res. 1995; 3(1): 51–58. https://doi.org/10.1016/0926-6410(95)00018-6 PMID: 8719022

129. Chokron S, Bartolomeo P, Perenin MT, Helft G, Imbert M. Scanning direction and line bisection: a study of normal subjects and unilateral neglect patients with opposite reading habits. Cogn Brain Res. 1998; 7(2): 173–178. https://doi.org/10.1016/s0926-6410(98)00022-6 PMID: 9774725

130. Sandson J, Albert ML. Varieties of perseveration. Neuropsychologia. 1984; 22(6): 715–732. https://doi.org/10.1016/0028-3932(84)90098-8 PMID: 6084826

131. Sandson J, Albert ML. Perseveration in behavioral neurology. Neurology. 1987; 37(11): 1736–1736. https://doi.org/10.1212/wnl.37.11.1736 PMID: 3670611

132. Hotz G, Helm-Estabrooks N. Perseveration. Part I: a review. Brain Inj. 1995; 9(2): 151–159. https://doi.org/10.3109/02699059509008188 PMID: 7787835

133. Ramage A, Bayles K, Helm-Estabrooks N, Cruz R. Frequency of perseveration in normal subjects. Brain Lang. 1999; 66(3): 329–340. https://doi.org/10.1006/brln.1999.2032 PMID: 10190994

134. Kimchi EY, Laubach M. The dorsomedial striatum reflects response bias during learning. J Neurosci. 2009; 29(47): 14891–14902. https://doi.org/10.1523/JNEUROSCI.4060-09.2009 PMID: 19940185

135. Banavar NV, Bornstein AM. Multi-plasticities: distinguishing context-specific habits from complex perseverations. In: Vandaele Y, editor. Habits: their definition, neurobiology and role in addiction. Cham, Switzerland: Springer Nature; 2024.

136. Thorndike EL. Animal intelligence: an experimental study of the associative processes in animals. Psychol Rev Monogr Suppl. 1898; 2(4): 1–109. https://doi.org/10.1037/h0092987

137. Thorndike EL. Animal intelligence: experimental studies. New York (NY): Macmillan; 1911. https://doi.org/10.5962/bhl.title.55072

138. Dickinson A. Actions and habits: the development of behavioural autonomy. Philos Trans R Soc Lond B Biol Sci. 1985; 308(1135): 67–78. https://doi.org/10.1098/rstb.1985.0010

139. Dayan P, Balleine BW. Reward, motivation, and reinforcement learning. Neuron. 2002; 36(2): 285–298. https://doi.org/10.1016/s0896-6273(02)00963-7 PMID: 12383782

140. Balleine BW O'Doherty JP. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology. 2010; 35: 48–69. https://doi.org/10.1038/npp.2009.131

141. Lally P, Van Jaarsveld CH, Potts HW, Wardle J. How are habits formed: modelling habit formation in the real world. Eur J Soc Psychol. 2010; 40(6): 998–1009. https://doi.org/10.1002/ejsp.674

142. Wood W, Rünger D. Psychology of habit. Annu Rev Psychol. 2016; 67: 289–314. https://doi.org/10.1146/annurev-psych-122414-033417 PMID: 26361052

143. Miller KJ, Shenhav A, Ludvig EA. Habits without values. Psychol Rev. 2019; 126(2): 292–311. https://doi.org/10.1037/rev0000120 PMID: 30676040

144. Bogacz R. Dopamine role in learning and action inference. eLife. 2020; 9: e53262. https://doi.org/10.7554/eLife.53262 PMID: 32633715

145. Richman CL, Dember WN, Kim P. Spontaneous alternation behavior in animals: a review. Curr Psychol Res Rev. 1986; 5(4): 358–391. https://doi.org/10.1007/bf02686603

146. Lalonde R. The neurobiological basis of spontaneous alternation. Neurosci Biobehav Rev. 2002; 26 (1): 91–104. https://doi.org/10.1016/s0149-7634(01)00041-0 PMID: 11835987

147. Pape AA, Siegel M. Motor cortex activity predicts response alternation during sensorimotor decisions. Nat Commun. 2016; 7(1): 1–10. https://doi.org/10.1038/ncomms13098 PMID: 27713396

148. Pape AA, Noury N, Siegel M. Motor actions influence subsequent sensorimotor decisions. Sci Rep. 2017; 7(1): 1–5. https://doi.org/10.1038/s41598-017-16299-0

149. Logan GD. On the ability to inhibit simple thoughts and actions: II. Stop-signal studies of repetition priming. J Exp Psychol Learn Mem Cogn. 1985; 11(4): 675–691. https://doi.org/10.1037/0278-7393.11.1-4.675

150. Jax SA, Rosenbaum DA. Hand path priming in manual obstacle avoidance: rapid decay of dorsal stream information. Neuropsychologia. 2009; 47(6): 1573–1577. https://doi.org/10.1016/j.neuropsychologia.2008.05.019 PMID: 18597796

151. Dixon P, McAnsh S, Read L. Repetition effects in grasping. Can J Exp Psychol. 2012; 66(1): 1–17. https://doi.org/10.1037/a0026192 PMID: 22148902

152. Glover S, Dixon P. Perseveration effects in reaching and grasping rely on motor priming and not perception. Exp Brain Res. 2013; 226: 53–61. https://doi.org/10.1007/s00221-013-3410-y PMID: 23354666

153. Valyear KF, Frey SH. Hand selection for object grasping is influenced by recent motor history. Psychon Bull Rev. 2014; 21: 566–573. https://doi.org/10.3758/s13423-013-0504-2 PMID: 24002968

154. Randerath J, Valyear KF, Hood A, Frey SH. Two routes to the same action: an action repetition priming study. J Mot Behav. 2015; 47(2): 142–152. https://doi.org/10.1080/00222895.2014.961891 PMID: 25350603

155. Valyear KF, Fitzpatrick AM, Dundon NM. Now and then: hand choice is influenced by recent action history. Psychon Bull Rev. 2019; 26: 305–314. https://doi.org/10.3758/s13423-018-1510-1 PMID: 30039397

156. Desimone R. Neural mechanisms for visual memory and their role in attention. Proc Natl Acad Sci U S A. 1996; 93(24): 13494–13499. https://doi.org/10.1073/pnas.93.24.13494 PMID: 8942962

157. Grill-Spector K, Henson R, Martin A. Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn Sci. 2006; 10(1): 14–23. https://doi.org/10.1016/j.tics.2005.11.006 PMID: 16321563

158. Grafton ST, Hamilton AF. Evidence for a distributed hierarchy of action representation in the brain. Hum Mov Sci. 2007; 26(4): 590–616. https://doi.org/10.1016/j.humov.2007.05.009 PMID: 17706312

159. Hamilton AF, Grafton ST. Repetition suppression for performed hand gestures revealed by fMRI. Hum Brain Mapp. 2009; 30(9): 2898–2906. https://doi.org/10.1002/hbm.20717 PMID: 19117276

160. Majdandžić J, Bekkering H, van Schie HT, Toni I. Movement-specific repetition suppression in ventral and dorsal premotor cortex during action observation. Cereb Cortex. 2009; 19(11): 2736–2745. https://doi.org/10.1093/cercor/bhp049 PMID: 19321652

161. Jurkiewicz MT, Gaetz WC, Bostan AC, Cheyne D. Post-movement beta rebound is generated in motor cortex: evidence from neuromagnetic recordings. NeuroImage. 2006; 32(3): 1281–1289. https://doi.org/10.1016/j.neuroimage.2006.06.005 PMID: 16863693

162. Briand KA, Larrison AL, Sereno AB. Inhibition of return in manual and saccadic response systems. Percept Psychophys. 2000; 62(8): 1512–1524. https://doi.org/10.3758/bf03212152 PMID: 11140175

163. Fecteau JH, Munoz DP. Exploring the consequences of the previous trial. Nat Rev Neurosci. 2003; 4 (6): 435–443. https://doi.org/10.1038/nrn1114 PMID: 12778116

164. Pastötter B, Hanslmayr S, Bäuml KH. Inhibition of return arises from inhibition of response processes: an analysis of oscillatory beta activity. J Cogn Neurosci. 2008; 20(1): 65–75. https://doi.org/10.1162/jocn.2008.20010 PMID: 17919085

165. Tune GS. A brief survey of variables that influence random-generation. Percept Mot Skills. 1964; 18 (3): 705–710. https://doi.org/10.2466/pms.1964.18.3.705 PMID: 14172516

166. Baddeley AD. The capacity for generating information by randomization. Q J Exp Psychol. 1966; 18 (2): 119–129. https://doi.org/10.1080/14640746608400019 PMID: 5935121

167. Wagenaar WA. Generation of random sequences by human subjects: a critical survey of literature. Psychol Bull. 1972; 77(1): 65–72. https://doi.org/10.1037/h0032060

168. Lopes LL. Doing the impossible: a note on induction and the experience of randomness. J Exp Psychol Learn Mem Cogn. 1982; 8(6): 626–636. https://doi.org/10.1037/0278-7393.8.6.626

169. Wiegersma S. Sequential response bias in randomized response sequences: a computer simulation. Acta Psychol. 1982; 52(3): 249–256. https://doi.org/10.1016/0001-6918(82)90011-7

170. Kareev Y. Not that bad after all: generation of random sequences. J Exp Psychol Hum Percept Perform. 1992; 18(4): 1189–1194. https://doi.org/10.1037/0096-1523.18.4.1189

171. Nickerson RS. The production and perception of randomness. Psychol Rev. 2002; 109(2): 330–357. https://doi.org/10.1037/0033-295x.109.2.330 PMID: 11990321

172. Lages M, Jaworska K. How predictable are "spontaneous decisions" and "hidden intentions"? Comparing classification results based on previous responses with multivariate pattern analysis of fMRI BOLD signals. Front Psychol. 2012; 3: 56. https://doi.org/10.3389/fpsyg.2012.00056 PMID: 22408630

173. Allefeld C, Soon CS, Bogler C, Heinzle J, Haynes JD. Sequential dependencies between trials in free choice tasks. arXiv. 2013; 1311.0753. https://doi.org/10.48550/arxiv.1311.0753

174. Guseva M, Bogler C, Allefeld C, Haynes JD. Instruction effects on randomness in sequence generation. Front Psychol. 2023; 14: 1113654. https://doi.org/10.3389/fpsyg.2023.1113654 PMID: 37034908

175. Castillo L, León-Villagrá P, Chater N, Sanborn A. Explaining the flaws in human random generation as local sampling with momentum. PLOS Comput Biol. 2024; 20(1): e1011739. https://doi.org/10.1371/journal.pcbi.1011739 PMID: 38181041

176. Parush N, Tishby N, Bergman H. Dopaminergic balance between reward maximization and policy complexity. Front Syst Neurosci. 2011; 5: 22. https://doi.org/10.3389/fnsys.2011.00022 PMID: 21603228

177. den Ouden HE, Daw ND, Fernandez G, Elshout JA, Rijpkema M, Hoogman M, Franke B, Cools R. Dissociable effects of dopamine and serotonin on reversal learning. Neuron. 2013; 80(4): 1090–1100. https://doi.org/10.1016/j.neuron.2013.08.030 PMID: 24267657

178. Greenstreet F, Vergara HM, Pati S, Schwarz L, Wisdom M, Marbach F, Johansson Y, Rollik L, Moskovitz T, Clopath C, Stephenson-Jones M. Action prediction error: a value-free dopaminergic teaching signal that drives stable learning. bioRxiv. 2022; 507572. https://doi.org/10.1101/2022.09.12.507572

179. Bari BA, Gershman SJ. Undermatching is a consequence of policy compression. J Neurosci. 2023; 43(3): 447–457. https://doi.org/10.1523/JNEUROSCI.1003-22.2022 PMID: 36639891

180. Grill F, Guitart-Masip M, Johansson J, Stiernman L, Axelsson J, Nyberg L, Rieckmann A. Dopamine release in human associative striatum during reversal learning. Nat Commun. 2024; 15: 59. https://doi.org/10.1038/s41467-023-44358-w PMID: 38167691

181. Ihara K, Shikano Y, Kato S, Yagishita S, Tanaka KF, Takata N. A reinforcement learning model with choice traces for a progressive ratio schedule. Front Behav Neurosci. 2024; 17: 1302842. https://doi.org/10.3389/fnbeh.2023.1302842 PMID: 38268795

182. Chakroun K, Mathar D, Wiehler A, Ganzer F, Peters J. Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. eLife. 2020; 9: e51260. https://doi.org/10.7554/eLife.51260 PMID: 32484779

183. Seymour B, Daw ND, Roiser JP, Dayan P, Dolan R. Serotonin selectively modulates reward value in human decision-making. J Neurosci. 2012; 32(17): 5833–5842. https://doi.org/10.1523/JNEUROSCI.0053-12.2012 PMID: 22539845

184. Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 1996; 16(5): 1936–1947. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996 PMID: 8774460

185. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275(5306): 1593–1599. https://doi.org/10.1126/science.275.5306.1593 PMID: 9054347

186. Schultz W. Neuronal reward and decision signals: from theories to data. Physiol Rev. 2015; 95(3): 853–951. https://doi.org/10.1152/physrev.00023.2014 PMID: 26109341

187. Kurniawan IT, Guitart-Masip M, Dolan RJ. Dopamine and effort-based decision making. Front Neurosci. 2011; 5: 81. https://doi.org/10.3389/fnins.2011.00081 PMID: 21734862

188. Barter JW, Li S, Lu D, Bartholomew RA, Rossi MA, Shoemaker CT, Salas-Meza D, Gaidis E, Yin HH. Beyond reward prediction errors: the role of dopamine in movement kinematics. Front Integr Neurosci. 2015; 9: 39. https://doi.org/10.3389/fnint.2015.00039 PMID: 26074791

189. Graybiel AM, Grafton ST. The striatum: where skills and habits meet. Cold Spring Harb Perspect Biol. 2015; 7(8): a021691. https://doi.org/10.1101/cshperspect.a021691 PMID: 26238359

190. Panigrahi B, Martin KA, Li Y, Graves AR, Vollmer A, Olson L, Mensh BD, Karpova AY, Dudman JT. Dopamine is required for the neural representation and control of movement vigor. Cell. 2015; 162(6): 1418–1430. https://doi.org/10.1016/j.cell.2015.08.014 PMID: 26359992

191. Walton ME, Bouret S. What is the relationship between dopamine and effort? Trends Neurosci. 2019; 42(2): 79–91. https://doi.org/10.1016/j.tins.2018.10.001 PMID: 30391016

192. Bakhurin K, Hughes RN, Jiang Q, Hossain M, Gutkin B, Fallon IP, Yin HH. Force tuning explains changes in phasic dopamine signaling during stimulus-reward learning. bioRxiv. 2023: 537994. https://doi.org/10.1101/2023.04.23.537994 PMID: 37162997

193. Brehm JW. Postdecision changes in the desirability of alternatives. J Abnorm Soc Psychol. 1956; 52 (3): 384–389. https://doi.org/10.1037/h0041006 PMID: 13318848

194. Festinger L. A theory of cognitive dissonance. Stanford (CA): Stanford University Press; 1957.

195. Izuma K, Matsumoto M, Murayama K, Samejima K, Sadato N, Matsumoto K. Neural correlates of cognitive dissonance and choice-induced preference change. Proc Natl Acad Sci U S A. 2010; 107(51): 22014–22019. https://doi.org/10.1073/pnas.1011879108 PMID: 21135218

196. Nakao T, Ohira H, Northoff G. Distinction between externally vs. internally guided decision-making: operational differences, meta-analytical comparisons and their theoretical implications. Front Neurosci. 2012; 6: 31. https://doi.org/10.3389/fnins.2012.00031 PMID: 22403525

197. Izuma K, Murayama K. Choice-induced preference change in the free-choice paradigm: a critical methodological review. Front Psychol. 2013; 4: 41. https://doi.org/10.3389/fpsyg.2013.00041 PMID: 23404185

198. Nakao T, Kanayama N, Katahira K, Odani M, Ito Y, Hirata Y, Nasuno R, Ozaki H, Hiramoto R, Miyatani M, Northoff G. Post-response βγ power predicts the degree of choice-based learning in internally guided decision-making. Sci Rep. 2016; 6: 32477. https://doi.org/10.1038/srep32477

199. Zhu J, Hashimoto J, Katahira K, Hirakawa M, Nakao T. Computational modeling of choice-induced preference change: a reinforcement-learning-based approach. PLOS ONE. 2021; 16(1): e0244434. https://doi.org/10.1371/journal.pone.0244434 PMID: 33411720

200. Toyama A, Katahira K, Ohira H. Reinforcement learning with parsimonious computation and a forgetting process. Front Hum Neurosci. 2019; 13: 153. https://doi.org/10.3389/fnhum.2019.00153 PMID: 31143107

201. Akam T, Rodrigues-Vaz I, Marcelo I, Zhang X, Pereira M, Oliveira RF, Dayan P, Costa RM. The anterior cingulate cortex predicts future states to mediate model-based action selection. Neuron. 2021; 109(1): 149–163. https://doi.org/10.1016/j.neuron.2020.10.013 PMID: 33152266

202. Rmus M, Zou A, Collins AG. Choice type impacts human reinforcement learning. J Cogn Neurosci. 2023; 35(2): 314–330. https://doi.org/10.1162/jocn_a_01947

203. Karagoz AB, Reagh ZM, Kool W. The construction and use of cognitive maps in model-based control. J Exp Psychol Gen. 2024; 153(2): 372–385. https://doi.org/10.1037/xge0001491 PMID: 38059968

204. Bouchacourt F, Palminteri S, Koechlin E, Ostojic S. Temporal chunking as a mechanism for unsupervised learning of task-sets. eLife. 2020; 9: e50469. https://doi.org/10.7554/eLife.50469 PMID: 32149602

205. Lai L, Gershman SJ. Policy compression: an information bottleneck in action selection. In: Federmeier KD, editor. The psychology of learning and motivation: vol. 74. Cambridge (MA): Academic Press; 2021. p. 195–232. https://doi.org/10.1016/s0079-7421(21)x0002-3

206. Lai L, Huang AZ, Gershman SJ. Action chunking as policy compression. PsyArXiv. 2022; z8yrv. https://doi.org/10.31234/osf.io/z8yrv

207. Akaishi R, Umeda K, Nagase A, Sakai K. Autonomous mechanism of internal choice estimate underlies decision inertia. Neuron. 2014; 81(1): 195–206. https://doi.org/10.1016/j.neuron.2013.10.018 PMID: 24333055

208. Thiel SD, Bitzer S, Nierhaus T, Kalberlah C, Preusser S, Neumann J, Nikulin VV, van der Meer E, Villringer A, Pleger B. Hysteresis as an implicit prior in tactile spatial decision making. PLOS ONE. 2014; 9(2): e89802. https://doi.org/10.1371/journal.pone.0089802 PMID: 24587045

209. Kaneko Y, Sakai K. Dissociation in decision bias mechanism between probabilistic information and previous decision. Front Hum Neurosci. 2015; 9: 261. https://doi.org/10.3389/fnhum.2015.00261 PMID: 25999844

210. Abrahamyan A, Silva LL, Dakin SC, Carandini M, Gardner JL. Adaptable history biases in human perceptual decisions. Proc Natl Acad Sci U S A. 2016; 113(25): E3548–E3557. https://doi.org/10.1073/pnas.1518786113 PMID: 27330086

211. Fritsche M, Mostert P, de Lange FP. Opposite effects of recent history on perception and decision. Curr Biol. 2017; 27(4): 590–595. https://doi.org/10.1016/j.cub.2017.01.006 PMID: 28162897

212. Braun A, Urai AE, Donner TH. Adaptive history biases result from confidence-weighted accumulation of past choices. J Neurosci. 2018; 38(10): 2418–2429. https://doi.org/10.1523/JNEUROSCI.2189-17.2017 PMID: 29371318

213. Schlunegger D, Mast FW. Probabilistic integration of preceding responses explains response bias in perceptual decision making. iScience. 2023; 26: 107123. https://doi.org/10.1016/j.isci.2023.107123 PMID: 37434696

214. Padoa-Schioppa C. Neuronal origins of choice variability in economic decisions. Neuron. 2013; 80(5): 1322–1336. https://doi.org/10.1016/j.neuron.2013.09.013 PMID: 24314733

215. Scherbaum S, Frisch S, Leiberg S, Lade SJ, Goschke T, Dshemuchadse M. Process dynamics in delay discounting decisions: an attractor dynamics approach. Judgm Decis Mak. 2016; 11(5): 472–495. https://doi.org/10.1017/S1930297500004575

216. Schoemann M, Scherbaum S. Choice history bias in intertemporal choice. PsyArXiv. 2020; 7h9zj. https://doi.org/10.31234/osf.io/7h9zj

217. Banavar NV, Bornstein AM. Independent, not irrelevant: trial order causes systematic misestimation of economic choice traits. PsyArXiv. 2023; a8gz3. https://doi.org/10.31234/osf.io/a8gz3

218. Bertelson P. Serial choice reaction-time as a function of response versus signal-and-response repetition. Nature. 1965; 206(980): 217–218. https://doi.org/10.1038/206217a0 PMID: 5830165

219. Pashler H, Baylis GC. Procedural learning: II. Intertrial repetition effects in speeded-choice tasks. J Exp Psychol Learn Mem Cogn. 1991; 17(1): 33–48. https://doi.org/10.1037/0278-7393.17.1.33

220. Cho RY, Nystrom LE, Brown ET, Jones AD, Braver TS, Holmes PJ, Cohen JD. Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task. Cogn Affect Behav Neurosci. 2002; 2(4): 283–299. https://doi.org/10.3758/cabn.2.4.283 PMID: 12641174

221. Fründ I, Wichmann FA, Macke JH. Quantifying the effect of intertrial dependence on perceptual decisions. J Vis. 2014; 14(7): 9. https://doi.org/10.1167/14.7.9 PMID: 24944238

222. Hwang EJ, Dahlen JE, Mukundan M, Komiyama T. History-based action selection bias in posterior parietal cortex. Nat Commun. 2017; 8(1): 1–14. https://doi.org/10.1038/s41467-017-01356-z

223. Akrami A, Kopec CD, Diamond ME, Brody CD. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. Nature. 2018; 554(7692): 368–372. https://doi.org/10.1038/nature25510 PMID: 29414944

224. Bosch E, Fritsche M, Ehinger BV, de Lange FP. Opposite effects of choice history and evidence history resolve a paradox of sequential choice bias. J Vis. 2020; 20(12): 9. https://doi.org/10.1167/jov.20.12.9 PMID: 33211062

225. Senftleben U, Schoemann M, Scherbaum S. Choice repetition bias in intertemporal choice: an eye-tracking study. PsyArXiv. 2024; g3v9m. https://doi.org/10.31234/osf.io/g3v9m

226. Gibson JJ. The ecological approach to visual perception. Boston (MA): Houghton Mifflin; 1979. https://doi.org/10.4324/9781315740218

227. Cisek P. Cortical mechanisms of action selection: the affordance competition hypothesis. Philos Trans R Soc Lond B Biol Sci. 2007; 362(1485): 1585–1599. https://doi.org/10.1098/rstb.2007.2054 PMID: 17428779

228. Cisek P, Kalaska JF. Neural mechanisms for interacting with a world full of action choices. Annu Rev Neurosci. 2010; 33: 269–298. https://doi.org/10.1146/annurev.neuro.051508.135409 PMID: 20345247

229. Cisek P. Making decisions through a distributed consensus. Curr Opin Neurobiol. 2012; 22(6): 927–936. https://doi.org/10.1016/j.conb.2012.05.007 PMID: 22683275

230. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011; 69(6): 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027 PMID: 21435563

231. Wimmer GE, Daw ND, Shohamy D. Generalization of value in reinforcement learning by humans. Eur J Neurosci. 2012; 35(7): 1092–1104. https://doi.org/10.1111/j.1460-9568.2012.08017.x PMID: 22487039

232. Christakou A, Gershman SJ, Niv Y, Simmons A, Brammer M, Rubia K. Neural and psychological maturation of decision-making in adolescence and young adulthood. J Cogn Neurosci. 2013; 25(11): 1807–1823. https://doi.org/10.1162/jocn_a_00447 PMID: 23859647

233. Voon V, Derbyshire K, Rück C, Irvine MA, Worbe Y, Enander J, Schreiber L, Gillan C, Fineberg NA, Sahakian BJ, Robbins TW, Harrison NA, Wood J, Daw ND, Dayan P, Grant JE, Bullmore ET. Disorders of compulsivity: a common bias towards learning habits. Mol Psychiatry. 2014; 20(3): 345–352. https://doi.org/10.1038/mp.2014.44 PMID: 24840709

234. Wimmer GE, Braun EK, Daw ND, Shohamy D. Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. J Neurosci. 2014; 34(45): 14901–14912. https://doi.org/10.1523/JNEUROSCI.0204-14.2014 PMID: 25378157

235. Balcarras M, Ardid S, Kaping D, Everling S, Womelsdorf T. Attentional selection can be predicted by reinforcement learning of task-relevant stimulus features weighted by value-independent stickiness. J Cogn Neurosci. 2016; 28(2): 333–349. https://doi.org/10.1162/jocn_a_00894 PMID: 26488586

236. Kool W, Cushman FA, Gershman SJ. When does model-based control pay off? PLOS Comput Biol. 2016; 12(8): e1005090. https://doi.org/10.1371/journal.pcbi.1005090 PMID: 27564094

237. Kool W, Gershman SJ, Cushman FA. Cost-benefit arbitration between multiple reinforcement-learning systems. Psychol Sci. 2017; 28(9): 1321–1333. https://doi.org/10.1177/0956797617708288 PMID: 28731839

238. Correa CM, Noorman S, Jiang J, Palminteri S, Cohen MX, Lebreton M, van Gaal S. How the level of reward awareness changes the computational and electrophysiological signatures of reinforcement learning. J Neurosci. 2018; 38(48): 10338–10348. https://doi.org/10.1523/JNEUROSCI.0457-18.2018 PMID: 30327418

239. Bruckner R, Nassar MR, Li SC, Eppinger B. Differences in learning across the lifespan emerge via resource-rational computations. PsyArXiv. 2020; nh9bq. https://doi.org/10.31234/osf.io/nh9bq

240. Miranda B, Malalasekera WN, Behrens TE, Dayan P, Kennerley SW. Combined model-free and model-sensitive reinforcement learning in non-human primates. PLOS Comput Biol. 2020; 16(6): e1007944. https://doi.org/10.1371/journal.pcbi.1007944 PMID: 32569311

241. Gueguen MC, Lopez-Persem A, Billeke P, Lachaux JP, Rheims S, Kahane P, Minotti L, David O, Pessiglione M, Bastin J. Anatomical dissociation of intracerebral signals for reward and punishment prediction errors in humans. Nat Commun. 2021; 12(1): 3344. https://doi.org/10.1038/s41467-021-23704-w PMID: 34099678

242. Eckstein MK, Master SL, Xia L, Dahl RE, Wilbrecht L, Collins AG. The interpretation of computational model parameters depends on the context. eLife. 2022; 11: e75474. https://doi.org/10.7554/eLife.75474 PMID: 36331872

243. Kovach CK, Daw ND, Rudrauf D, Tranel D, O'Doherty JP, Adolphs R. Anterior prefrontal cortex contributes to action selection through tracking of recent reward trends. J Neurosci. 2012; 32(25): 8434–8442. https://doi.org/10.1523/JNEUROSCI.5468-11.2012 PMID: 22723683

244. Haines N, Vassileva J, Ahn WY. The outcome-representation learning model: a novel reinforcement learning model of the Iowa gambling task. Cogn Sci. 2018; 42(8): 2534–2561. https://doi.org/10.1111/cogs.12688 PMID: 30289167

245. Iigaya K, Fonseca MS, Murakami M, Mainen ZF, Dayan P. An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. Nat Commun. 2018; 9: 2477. https://doi.org/10.1038/s41467-018-04840-2 PMID: 29946069

246. Ebitz RB, Sleezer BJ, Jedema HP, Bradberry CW, Hayden BY. Tonic exploration governs both flexibility and lapses. PLOS Comput Biol. 2019; 15(11): e1007475. https://doi.org/10.1371/journal.pcbi.1007475 PMID: 31703063

247. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553): 436–444. https://doi.org/10.1038/nature14539 PMID: 26017442

248. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015; 61: 85–117. https://doi.org/10.1016/j.neunet.2014.09.003 PMID: 25462637

249. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge (MA): MIT Press; 2016.

250. Barak O. Recurrent neural networks as versatile tools of neuroscience research. Curr Opin Neurobiol. 2017; 46: 1–6. https://doi.org/10.1016/j.conb.2017.06.003 PMID: 28668365

251. Ma WJ, Peters B. A neural network walks into a lab: towards using deep nets as models for human behavior. arXiv. 2020; 2005.02181. https://doi.org/10.48550/arXiv.2005.02181

252. Tesauro G. Temporal difference learning and TD-Gammon. Commun ACM. 1995; 38(3): 58–68. https://doi.org/10.1145/203330.203343

253. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep reinforcement learning. Nature. 2015; 518(7540): 529–533. https://doi.org/10.1038/nature14236 PMID: 25719670

254. Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA. Deep reinforcement learning: a brief survey. IEEE Signal Process Mag. 2017; 34(6): 26–38. https://doi.org/10.1109/msp.2017.2743240

255. Li Y. Deep reinforcement learning: an overview. arXiv. 2017; 1701.07274. https://doi.org/10.48550/arxiv.1701.07274

256. Sünderhauf N, Brock O, Scheirer W, Hadsell R, Fox D, Leitner J, Upcroft B, Abbeel P, Burgard W, Milford M, Corke P. The limits and potentials of deep learning for robotics. Int J Rob Res. 2018; 37(4–5): 405–420. https://doi.org/10.1177/0278364918770733

257. Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Blundell C, Hassabis D. Reinforcement learning, fast and slow. Trends Cogn Sci. 2019; 23(5): 408–422. https://doi.org/10.1016/j.tics.2019.02.006 PMID: 31003893

258. Nguyen H, La H. Review of deep reinforcement learning for robot manipulation. IEEE Int Conf Robot Comput. 2019; 3: 590–595. https://doi.org/10.1109/irc.2019.00120

259. Botvinick M, Wang JX, Dabney W, Miller KJ, Kurth-Nelson Z. Deep reinforcement learning and its neuroscientific implications. Neuron. 2020; 107(4): 603–616. https://doi.org/10.1016/j.neuron.2020.06.014 PMID: 32663439

260. Ibarz J, Tan J, Finn C, Kalakrishnan M, Pastor P, Levine S. How to train your robot with deep reinforcement learning: lessons we have learned. Int J Rob Res. 2021; 40(4–5): 698–721. https://doi.org/10.1177/0278364920987859

261. Amari SI. Learning patterns and pattern sequences by self-organizing nets of threshold elements. IEEE Trans Comput. 1972; C-21(11): 1197–1206. https://doi.org/10.1109/t-c.1972.223477

262. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A. 1982; 79(8): 2554–2558. https://doi.org/10.1073/pnas.79.8.2554 PMID: 6953413

263. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986; 323(6088): 533–536. https://doi.org/10.1038/323533a0

264. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997; 9(8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 PMID: 9377276

265. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv. 2014; 1406.1078. https://doi.org/10.48550/arxiv.1406.1078

266. Song HF, Yang GR, Wang XJ. Reward-based training of recurrent neural networks for cognitive and value-based tasks. eLife. 2017; 6: e21492. https://doi.org/10.7554/eLife.21492 PMID: 28084991

267. Dezfouli A, Griffiths K, Ramos F, Dayan P, Balleine BW. Models that learn how humans learn: the case of decision-making and its disorders. PLOS Comput Biol. 2019; 15(6): e1006903. https://doi.org/10.1371/journal.pcbi.1006903 PMID: 31185008

268. Dezfouli A, Ashtiani H, Ghattas O, Nock R, Dayan P, Ong CS. Disentangled behavioural representations. Adv Neural Inf Process Syst. 2019; 32.

269. Kim D, Lee JH, Shin JH, Yang MA, Lee SW. On the reliability and generalizability of brain-inspired reinforcement learning algorithms. arXiv. 2020; 2007.04578. https://doi.org/10.48550/arxiv.2007.04578

270. Fintz M, Osadchy M, Hertz U. Using deep learning to predict human decisions and using cognitive models to explain deep learning models. Sci Rep. 2022; 12: 4736. https://doi.org/10.1038/s41598-022-08863-0 PMID: 35304572

271. Eckstein MK, Summerfield C, Daw ND, Miller KJ. Predictive and interpretable: combining artificial neural networks and classic cognitive models to understand human learning and decision making. bioRxiv. 2023; 541226. https://doi.org/10.1101/2023.05.17.541226

272. Kim D, Lee JH, Jung W, Kim SH, Lee SW. Long short-term prediction guides human metacognitive reinforcement learning. Res Sq. 2023; 3080402. https://doi.org/10.21203/rs.3.rs-3080402/v1

273. Kuperwajs I, Schütt HH, Ma WJ. Using deep neural networks as a guide for modeling human planning. Sci Rep. 2023; 13: 20269. https://doi.org/10.1038/s41598-023-46850-1 PMID: 37985896

274. Li J, Benna MK, Mattar MG. Automatic discovery of cognitive strategies with tiny recurrent neural networks. bioRxiv. 2023; 536629. https://doi.org/10.1101/2023.04.12.536629

275. Miller KJ, Eckstein M, Botvinick MM, Kurth-Nelson Z. Cognitive model discovery via disentangled RNNs. Adv Neural Inf Process Syst. 2023; 36.

276. Rmus M, Pan TF, Xia L, Collins AG. Artificial neural networks for model identification and parameter estimation in computational cognitive models. bioRxiv. 2023; 557793. https://doi.org/10.1101/2023.09.14.557793 PMID: 37767088

277. Tuzsus D, Pappas I, Peters J. Human-level reinforcement learning performance of recurrent neural networks is linked to hyperperseveration, not directed exploration. bioRxiv. 2023; 538570. https://doi.org/10.1101/2023.04.27.538570

278. Ger Y, Nachmani E, Wolf L, Shahar N. Harnessing the flexibility of neural networks to predict dynamic theoretical parameters underlying human choice behavior. PLOS Comput Biol. 2024; 20(1): e1011678. https://doi.org/10.1371/journal.pcbi.1011678 PMID: 38175848

279. Ger Y, Shahar M, Shahar N. Using recurrent neural network to estimate irreducible stochasticity in human choice-behavior. eLife. 2024; 13: e90082. https://doi.org/10.7554/elife.90082

280. Navarro DJ. Between the devil and the deep blue sea: tensions between scientific judgement and statistical model selection. Comput Brain Behav. 2019; 2: 28–34. https://doi.org/10.1007/s42113-018-0019-z

281. Karpathy A, Johnson J, Fei-Fei L. Visualizing and understanding recurrent networks. arXiv. 2015; 1506.02078. https://doi.org/10.48550/arxiv.1506.02078

282. Alharin A, Doan TN, Sartipi M. Reinforcement learning interpretation methods: a survey. IEEE Access. 2020; 8: 171058–171077. https://doi.org/10.1109/access.2020.3023394

283. Molnar C, Casalicchio G, Bischl B. Interpretable machine learning—a brief history, state-of-the-art and challenges. In: Koprinska I, Kamp M, Appice A, Loglisci C, Antonie L, Zimmermann A, Guidotti R, Özgöbek Ö, editors. Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020). Cham, Switzerland: Springer; 2020. p. 417–431. https://doi.org/10.1007/978-3-030-65965-3_28

284. Puiutta E, Veith EM. Explainable reinforcement learning: a survey. In: Holzinger A, Kieseberg P, Tjoa AM, Weippl E, editors. Machine Learning and Knowledge Extraction: International Cross-Domain Conference (CD-MAKE 2020). Cham, Switzerland: Springer; 2020. p. 77–95. https://doi.org/10.1007/978-3-030-57321-8_5

285. Glanois C, Weng P, Zimmer M, Li D, Yang T, Hao J, Liu W. A survey on interpretable reinforcement learning. arXiv. 2021; 2112.13112. https://doi.org/10.48550/arxiv.2112.13112

286. Heuillet A, Couthouis F, Díaz-Rodríguez N. Explainability in deep reinforcement learning. Knowl Based Syst. 2021; 214: 106685. https://doi.org/10.1016/j.knosys.2020.106685

287. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining deep neural networks and beyond: A review of methods and applications. Proc IEEE. 2021; 109(3): 247–278. https://doi.org/10.1109/jproc.2021.3060483

288. Akrour R, Tateo D, Peters J. Continuous action reinforcement learning from a mixture of interpretable experts. IEEE Trans Pattern Anal Mach Intell. 2022; 44(10): 6795–6806. https://doi.org/10.1109/TPAMI.2021.3103132 PMID: 34375280

289. Chen Z, Deng Y, Wu Y, Gu Q, Li Y. Towards understanding the mixture-of-experts layer in deep learning. Adv Neural Inf Process Syst. 2022; 35.

290. Milani S, Topin N, Veloso M, Fang F. A survey of explainable reinforcement learning. arXiv. 2022; 2202.08434. https://doi.org/10.48550/arxiv.2202.08434

291. Vasić M, Petrović A, Wang K, Nikolić M, Singh R, Khurshid S. MoËT: Mixture of Expert Trees and its application to verifiable reinforcement learning. Neural Netw. 2022; 151: 34–47. https://doi.org/10.1016/j.neunet.2022.03.022 PMID: 35381441

292. Räuker T, Ho A, Casper S, Hadfield-Menell D. Toward transparent AI: a survey on interpreting the inner structures of deep neural networks. In: 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). Raleigh (NC): IEEE; 2023. p. 464–483. https://doi.org/10.1109/SaTML54575.2023.00039

293. Cisek P, Puskas GA, El-Murr S. Decisions in changing conditions: the urgency-gating model. J Neurosci. 2009; 29(37): 11560–11571. https://doi.org/10.1523/JNEUROSCI.1844-09.2009 PMID: 19759303

294. Caie B, Endres D, Khan AZ, Blohm G. Choice anticipation as gated accumulation of sensory expectations. bioRxiv. 2023; 571751. https://doi.org/10.1101/2023.12.14.571751

295. Bertelson P. Sequential redundancy and speed in a serial two-choice responding task. Q J Exp Psychol. 1961; 13(2): 90–102. https://doi.org/10.1080/17470216108416478

296. Soetens E, Deboeck M, Hueting J. Automatic aftereffects in two-choice reaction time: a mathematical representation of some concepts. J Exp Psychol Hum Percept Perform. 1984; 10(4): 581–598. https://doi.org/10.1037//0096-1523.10.4.581 PMID: 6235321

297. Soetens E, Boer LC, Hueting JE. Expectancy or automatic facilitation? Separating sequential effects in two-choice reaction time. J Exp Psychol Hum Percept Perform. 1985; 11(5): 598–616. https://doi.org/10.1037/0096-1523.11.5.598

298. Rustichini A, Padoa-Schioppa C. A neuro-computational model of economic decisions. J Neurophysiol. 2015; 114(3): 1382–1398. https://doi.org/10.1152/jn.00184.2015 PMID: 26063776

299. Bonaiuto JJ, de Berker A, Bestmann S. Response repetition biases in human perceptual decisions are explained by activity decay in competitive attractor models. eLife. 2016; 5: e20047. https://doi.org/10.7554/eLife.20047 PMID: 28005007

300. Senftleben U, Schoemann M, Schwenke D, Richter S, Dshemuchadse M, Scherbaum S. Choice perseveration in value-based decision making: the impact of inter-trial interval and mood. Acta Psychol. 2019; 198: 102876. https://doi.org/10.1016/j.actpsy.2019.102876 PMID: 31280037

301. Senftleben U, Schoemann M, Rudolf M, Scherbaum S. To stay or not to stay: the stability of choice perseveration in value-based decision making. Q J Exp Psychol. 2021; 74(1): 199–217. https://doi.org/10.1177/1747021820964330 PMID: 32976065

302. Katahira K. How hierarchical models improve point estimates of model parameters at the individual level. J Math Psychol. 2016; 73: 37–58. https://doi.org/10.1016/j.jmp.2016.03.007

303. Ahn WY, Haines N, Zhang L. Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. Comput Psychiatr. 2017; 1: 24–57. https://doi.org/10.1162/CPSY_a_00002 PMID: 29601060

304. Piray P, Dezfouli A, Heskes T, Frank MJ, Daw ND. Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. PLOS Comput Biol. 2019; 15(6): e1007043. https://doi.org/10.1371/journal.pcbi.1007043 PMID: 31211783

305. van Geen C, Gerraty RT. Hierarchical Bayesian models of reinforcement learning: introduction and comparison to alternative methods. J Math Psychol. 2021; 105: 102602. https://doi.org/10.1016/j.jmp.2021.102602

306. Moutoussis M, Bullmore ET, Goodyer IM, Fonagy P, Jones PB, Dolan RJ, Dayan P, Neuroscience in Psychiatry Network Research Consortium. Change, stability, and instability in the Pavlovian guidance of behaviour from adolescence to young adulthood. PLOS Comput Biol. 2018; 14(12): e1006679. https://doi.org/10.1371/journal.pcbi.1006679

307. Enkavi AZ, Eisenberg IW, Bissett PG, Mazza GL, MacKinnon DP, Marsch LA, Poldrack RA. Large-scale analysis of test-retest reliabilities of self-regulation measures. Proc Natl Acad Sci U S A. 2019; 116(12): 5472–5477. https://doi.org/10.1073/pnas.1818430116 PMID: 30842284

308. Shahar N, Hauser TU, Moutoussis M, Moran R, Keramati M, NSPN Consortium, Dolan RJ. Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. PLOS Comput Biol. 2019; 15(2): e1006803. https://doi.org/10.1371/journal.pcbi.1006803 PMID: 30759077

309. Weidinger L, Gradassi A, Molleman L, van den Bos W. Test-retest reliability of canonical reinforcement learning models. Conf Cogn Comput Neurosci. 2019; 14: 513–516. https://doi.org/10.32470/ccn.2019.1053-0

310. Brown VM, Chen J, Gillan CM, Price RB. Improving the reliability of computational analyses: model-based planning and its relationship with compulsivity. Biol Psychiatry Cogn Neurosci Neuroimaging. 2020; 5(6): 601–609. https://doi.org/10.1016/j.bpsc.2019.12.019 PMID: 32249207

311. Haines N, Kvam PD, Irving LH, Smith C, Beauchaine TP, Pitt MA, Ahn W, Turner BM. Theoretically informed generative models can advance the psychological and brain sciences: lessons from the reliability paradox. PsyArXiv. 2020; xr7y3. https://doi.org/10.31234/osf.io/xr7y3

312. Eckstein MK, Wilbrecht L, Collins AG. What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. Curr Opin Behav Sci. 2021; 41: 128–137. https://doi.org/10.1016/j.cobeha.2021.06.004 PMID: 34984213

313. Pike AC, Tan K, Ansari HJ, Wing M, Robinson OJ. Test-retest reliability of affective bias tasks. PsyArXiv. 2022; n2fkh. https://doi.org/10.31234/osf.io/n2fkh

314. Sullivan-Toole H, Haines N, Dale K, Olino TM. Enhancing the psychometric properties of the Iowa gambling task using full generative modeling. Comput Psychiatr. 2022; 6(1): 189–212. https://doi.org/10.5334/cpsy.89 PMID: 37332395

315. Waltmann M, Schlagenhauf F, Deserno L. Sufficient reliability of the behavioral and computational readouts of a probabilistic reversal learning task. Behav Res Methods. 2022; 54(6): 2993–3014. https://doi.org/10.3758/s13428-021-01739-7 PMID: 35167111

316. Karvelis P, Paulus MP, Diaconescu AO. Individual differences in computational psychiatry: a review of current challenges. Neurosci Biobehav Rev. 2023; 148: 105137. https://doi.org/10.1016/j.neubiorev.2023.105137 PMID: 36940888

317. Mkrtchian A, Valton V, Roiser JP. Reliability of decision-making and reinforcement learning computational parameters. Comput Psychiatr. 2023; 7(1): 30–46. https://doi.org/10.5334/cpsy.86

318. Schaaf JV, Weidinger L, Molleman L, van den Bos W. Test-retest reliability of reinforcement learning parameters. PsyArXiv. 2023; chq5a. https://doi.org/10.3758/s13428-023-02203-4 PMID: 37684495

319. Schurr R, Reznik D, Hillman H, Bhui R, Gershman SJ. Dynamic computational phenotyping of human cognition. PsyArXiv. 2023; mgpqa. https://doi.org/10.31234/osf.io/mgpqa

320. Vrizzi S, Najar A, Lemogne C, Palminteri S, Lebreton M. Comparing the test-retest reliability of behavioral, computational and self-reported individual measures of reward and punishment sensitivity in relation to mental health symptoms. PsyArXiv. 2023; 3u4gp. https://doi.org/10.31234/osf.io/3u4gp

321. Efron B, Morris C. Stein's paradox in statistics. Sci Am. 1977; 236(5): 119–127. https://doi.org/10.1038/scientificamerican0577-119

322. Efron B. Empirical Bayes methods for combining likelihoods. J Am Stat Assoc. 1996; 91(434): 538–550. https://doi.org/10.2307/2291646

323. Huys QJ, Moutoussis M, Williams J. Are computational models of any use to psychiatry? Neural Netw. 2011; 24(6): 544–551. https://doi.org/10.1016/j.neunet.2011.03.001 PMID: 21459554

324. Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci. 2011; 14(2): 154–162. https://doi.org/10.1038/nn.2723 PMID: 21270784

325. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends Cogn Sci. 2012; 16 (1): 72–80. https://doi.org/10.1016/j.tics.2011.11.018 PMID: 22177032

326. Stephan KE, Schlagenhauf F, Huys QJ, Raman S, Aponte EA, Brodersen KH, Rigoux L, Moran RJ, Daunizeau J, Dolan RJ, Friston KJ, Heinz A. Computational neuroimaging strategies for single patient predictions. NeuroImage. 2017; 145: 180–199. https://doi.org/10.1016/j.neuroimage.2016.06.038 PMID: 27346545

327. Patzelt EH, Hartley CA, Gershman SJ. Computational phenotyping: using models to understand individual differences in personality, development, and mental illness. Personal Neurosci. 2018; 1: E18. https://doi.org/10.1017/pen.2018.14 PMID: 32435735

328. Haines N, Sullivan-Toole H, Olino T. From classical methods to generative models: tackling the unreliability of neuroscientific measures in mental health research. Biol Psychiatry Cogn Neurosci Neuroimaging. 2023; 8(8): 822–831. https://doi.org/10.1016/j.bpsc.2023.01.001 PMID: 36997406

329. O'Doherty JP, Hampton A, Kim H. Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci. 2007; 1104(1): 35–53. https://doi.org/10.1196/annals.1390.022 PMID: 17416921

330. Lebreton M, Bavard S, Daunizeau J, Palminteri S. Assessing inter-individual differences with task-related functional neuroimaging. Nat Hum Behav. 2019; 3(9): 897–905. https://doi.org/10.1038/s41562-019-0681-8 PMID: 31451737

331. Katahira K, Toyama A. Revisiting the importance of model fitting for model-based fMRI: it does matter in computational psychiatry. PLOS Comput Biol. 2021; 17(2): e1008738. https://doi.org/10.1371/journal.pcbi.1008738 PMID: 33561125

332. de Ruiter MB, Veltman DJ, Goudriaan AE, Oosterlaan J, Sjoerds Z, van den Brink W. Response perseveration and ventral prefrontal sensitivity to reward and punishment in male problem gamblers and smokers. Neuropsychopharmacology. 2009; 34(4): 1027–1038. https://doi.org/10.1038/npp.2008.175 PMID: 18830241

333. Gold JI, Law CT, Connolly P, Bennur S. The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. J Neurophysiol. 2008; 100(5): 2653–2668. https://doi.org/10.1152/jn.90629.2008 PMID: 18753326

334. Jones PR, Moore DR, Shub DE, Amitay S. The role of response bias in perceptual learning. J Exp Psychol Learn Mem Cogn. 2015; 41(5): 1456–1470. https://doi.org/10.1037/xlm0000111 PMID: 25867609

335. Urai AE, Braun A, Donner TH. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. Nat Commun. 2017; 8(1): 1–11. https://doi.org/10.1038/ncomms14637

336. Urai AE, de Gee JW, Tsetsos K, Donner TH. Choice history biases subsequent evidence accumulation. eLife. 2019; 8: e46331. https://doi.org/10.7554/eLife.46331 PMID: 31264959

337. Ratcliff R. A theory of memory retrieval. Psychol Rev. 1978; 85(2): 59–108. https://doi.org/10.1037/0033-295x.85.2.59

338. Busemeyer JR, Townsend JT. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychol Rev. 1993; 100(3): 432–459. https://doi.org/10.1037/0033-295x.100.3.432 PMID: 8356185

339. Usher M, McClelland JL. The time course of perceptual choice: the leaky, competing accumulator model. Psychol Rev. 2001; 108(3): 550–592. https://doi.org/10.1037/0033-295x.108.3.550 PMID: 11488378

340. Colas JT. Value-based decision making via sequential sampling with hierarchical competition and attentional modulation. PLOS ONE. 2017; 12(10): e0186822. https://doi.org/10.1371/journal.pone.0186822 PMID: 29077746

341. Wang ZJ, Busemeyer JR. Cognitive choice modeling. Cambridge (MA): MIT Press; 2021. https://doi.org/10.7551/mitpress/10469.001.0001

342. Garrett HE. A study of the relation of accuracy and speed. Arch Psychol. 1922; 56.

343. Johnson DM. Confidence and speed in the two-category judgment. Arch Psychol. 1939; 241.

344. Hull CL. Principles of behavior: an introduction to behavior theory. Oxford, United Kingdom: Appleton-Century-Crofts; 1943.

345. Kool W, McGuire JT, Rosen ZB, Botvinick MM. Decision making and the avoidance of cognitive demand. J Exp Psychol Gen. 2010; 139(4): 665–682. https://doi.org/10.1037/a0020198 PMID: 20853993

346. Dixon ML, Christoff K. The decision to engage cognitive control is driven by expected reward-value: neural and behavioral evidence. PLOS ONE. 2012; 7(12): e51637. https://doi.org/10.1371/journal.pone.0051637 PMID: 23284730

347. Shenhav A, Botvinick MM, Cohen JD. The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron. 2013; 79(2): 217–240. https://doi.org/10.1016/j.neuron.2013.07.007 PMID: 23889930

348. Westbrook A, Kester D, Braver TS. What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. PLOS ONE. 2013; 8(7): e68210. https://doi.org/10.1371/journal.pone.0068210 PMID: 23894295

349. Kool W, Botvinick M. A labor/leisure tradeoff in cognitive control. J Exp Psychol Gen. 2014; 43(1): 131–141. https://doi.org/10.1037/a0031048 PMID: 23230991

350. Botvinick M, Braver T. Motivation and cognitive control: from behavior to neural mechanism. Annu Rev Psychol. 2015; 66: 83–113. https://doi.org/10.1146/annurev-psych-010814-015044 PMID: 25251491

351. Westbrook A, Braver TS. Cognitive effort: A neuroeconomic approach. Cogn Affect Behav Neurosci. 2015; 15: 395–415. https://doi.org/10.3758/s13415-015-0334-y PMID: 25673005

352. Shenhav A, Cohen JD, Botvinick MM. Dorsal anterior cingulate cortex and the value of control. Nat Neurosci. 2016; 19(10): 1286–1291. https://doi.org/10.1038/nn.4384 PMID: 27669989

353. Kool W, Botvinick M. Mental labour. Nat Hum Behav. 2018; 2(12): 899–908. https://doi.org/10.1038/s41562-018-0401-9 PMID: 30988433

354. Pezzulo G, Rigoli F, Friston KJ. Hierarchical active inference: a theory of motivated control. Trends Cogn Sci. 2018; 22(4): 294–306. https://doi.org/10.1016/j.tics.2018.01.009 PMID: 29475638

355. Sidarus N, Palminteri S, Chambon V. Cost-benefit trade-offs in decision-making and learning. PLOS Comput Biol. 2019; 15(9): e1007326. https://doi.org/10.1371/journal.pcbi.1007326 PMID: 31490934

356. Zénon A, Solopchuk O, Pezzulo G. An information-theoretic perspective on the costs of cognition. Neuropsychologia. 2019; 123: 5–18. https://doi.org/10.1016/j.neuropsychologia.2018.09.013 PMID: 30268880

357. Gershman SJ. Origin of perseveration in the trade-off between reward and complexity. Cognition. 2020; 204: 104394. https://doi.org/10.1016/j.cognition.2020.104394 PMID: 32679270

358. Bhui R, Lai L, Gershman SJ. Resource-rational decision making. Curr Opin Behav Sci. 2021; 41: 15–21. https://doi.org/10.1016/j.cobeha.2021.02.015

359. Lai L, Gershman SJ. Human decision making balances reward maximization and policy compression. PsyArXiv. 2023; rnz72. https://doi.org/10.31234/osf.io/rnz72

360. Simon HA. Rational choice and the structure of the environment. Psychol Rev. 1956; 63(2): 129–138. https://doi.org/10.1037/h0042769 PMID: 13310708

361. Gigerenzer G, Brighton H. Homo heuristicus: why biased minds make better inferences. Top Cogn Sci. 2009; 1(1): 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x PMID: 25164802

362. Gigerenzer G, Gaissmaier W. Heuristic decision making. Annu Rev Psychol. 2011; 62: 451–482. https://doi.org/10.1146/annurev-psych-120709-145346 PMID: 21126183

363. Gao J, Wong-Lin K, Holmes P, Simen P, Cohen JD. Sequential effects in two-choice reaction time tasks: decomposition and synthesis of mechanisms. Neural Comput. 2009; 21(9): 2407–2436. https://doi.org/10.1162/neco.2009.09-08-866 PMID: 19548803

364. Tarantola T, Folke T, Boldt A, Pérez OD, Martino BD. Confirmation bias optimizes reward learning. bioRxiv. 2021; 433214. https://doi.org/10.1101/2021.02.27.433214

365. Lefebvre G, Summerfield C, Bogacz R. A normative account of confirmation bias during reinforcement learning. Neural Comput. 2022; 34(2): 307–337. https://doi.org/10.1162/neco_a_01455 PMID: 34758486

366. Fischer J, Whitney D. Serial dependence in visual perception. Nat Neurosci. 2014; 17(5): 738–743. https://doi.org/10.1038/nn.3689 PMID: 24686785

367. Ernst MR, Burwick T, Triesch J. Recurrent processing improves occluded object recognition and gives rise to perceptual hysteresis. J Vis. 2021; 21(13): 6. https://doi.org/10.1167/jov.21.13.6 PMID: 34905052

368. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science. 1974; 185 (4157): 1124–1131. https://doi.org/10.1126/science.185.4157.1124 PMID: 17835457

369. Lieder F, Griffiths TL, Huys QJ, Goodman ND. The anchoring bias reflects rational use of cognitive resources. Psychon Bull Rev. 2018; 25(1): 322–349. https://doi.org/10.3758/s13423-017-1286-8 PMID: 28484952

370. Lewin K. A dynamic theory of personality. New York (NY): McGraw-Hill; 1935.

371. Lewin K. Principles of topological psychology. New York (NY): McGraw-Hill; 1936. https://doi.org/10.1037/10019-000

372. Tolman EC. Cognitive maps in rats and men. Psychol Rev. 1948; 55(4): 189–208. https://doi.org/10.1037/h0061626 PMID: 18870876

373. Behrens TE, Muller TH, Whittington JC, Mark S, Baram AB, Stachenfeld KL, Kurth-Nelson Z. What is a cognitive map? Organizing knowledge for flexible behavior. Neuron. 2018; 100(2): 490–509. https://doi.org/10.1016/j.neuron.2018.10.002 PMID: 30359611

374. Joel D, Niv Y, Ruppin E. Actor-critic models of the basal ganglia: new anatomical and computational perspectives. Neural Netw. 2002; 15(4): 535–547. https://doi.org/10.1016/s0893-6080(02)00047-3 PMID: 12371510

375. Daw ND, Niv Y, Dayan P. Actions, values, policies, and the basal ganglia. In: Bezard E, editor. Recent breakthroughs in basal ganglia research. New York (NY): Nova Science; 2006a. p. 91–106.

376. Palminteri S, Boraud T, Lafargue G, Dubois B, Pessiglione M. Brain hemispheres selectively track the expected value of contralateral options. J Neurosci. 2009; 29(43): 13465–13472. https://doi.org/10.1523/JNEUROSCI.1500-09.2009 PMID: 19864559

377. Wunderlich K, Rangel A, O'Doherty JP. Neural computations underlying action-based decision making in the human brain. Proc Natl Acad Sci U S A. 2009; 106(40): 17199–17204. https://doi.org/10.1073/pnas.0901077106 PMID: 19805082

378. Giarrocco F, Costa VD, Basile BM, Pujara MS, Murray EA, Averbeck BB. Motor system-dependent effects of amygdala and ventral striatum lesions on explore-exploit behaviors. J Neurosci. 2023. https://doi.org/10.1523/jneurosci.1206-23.2023

379. Herrera D, Treviño M. Undesirable choice biases with small differences in the spatial structure of chance stimulus sequences. PLOS ONE. 2015; 10(8): e0136084. https://doi.org/10.1371/journal.pone.0136084 PMID: 26305097

380. Baldassarre G. A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviours. Cogn Syst Res. 2002; 3(1): 5–13. https://doi.org/10.1016/s1389-0417(01)00039-0

381. Khamassi M, Lachèze L, Girard B, Berthoz A, Guillot A. Actor-critic models of reinforcement learning in the basal ganglia: from natural to artificial rats. Adapt Behav. 2005; 13(2): 131–148. https://doi.org/10.1177/105971230501300205

382. Lee SW, Shimojo S, O'Doherty JP. Neural computations underlying arbitration between model-based and model-free learning. Neuron. 2014; 81(3): 687–699. https://doi.org/10.1016/j.neuron.2013.11.028 PMID: 24507199

383. Jordan MI, Jacobs RA. Hierarchical mixtures of experts and the EM algorithm. Neural Comput. 1994; 6(2): 181–214. https://doi.org/10.1162/neco.1994.6.2.181

384. Uchibe E, Doya K. Competitive-cooperative-concurrent reinforcement learning with importance sampling. In: Schaal S, Ijspeert AJ, Billard A, Vijayakumar S, Meyer J, editors. From Animals to Animats 8: Proceedings of the Eighth International Conference on the Simulation of Adaptive Behavior. Cambridge, MA: MIT Press; 2004. p. 287–296. https://doi.org/10.7551/mitpress/3122.003.0037

385. Bengio Y. Deep learning of representations: looking forward. In Dediu AH, Martín-Vide C, Mitkov R, Truthe B, editors. International Conference on Statistical Language and Speech Processing (SLSP 2013). Berlin, Germany: Springer; 2013. p. 1–37. https://doi.org/10.1007/978-3-642-39593-2_1

386. Bengio Y, Léonard N, Courville A. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv. 2013; 1308.3432. https://doi.org/10.48550/arxiv.1308.3432

387. Eigen D, Ranzato MA, Sutskever I. Learning factored representations in a deep mixture of experts. arXiv. 2013; 1312.4314. https://doi.org/10.48550/arxiv.1312.4314

388. Bengio E, Bacon PL, Pineau J, Precup D. Conditional computation in neural networks for faster models. arXiv. 2015; 1511.06297. https://doi.org/10.48550/arxiv.1511.06297

389. Graves A. Adaptive computation time for recurrent neural networks. arXiv. 2016; 1603.08983. https://doi.org/10.48550/arxiv.1603.08983

390. Peng XB, Berseth G, Van de Panne M. Terrain-adaptive locomotion skills using deep reinforcement learning. ACM Trans Graph. 2016; 35(4): 81. https://doi.org/10.1145/2897824.2925881

391. Peng XB, Chang M, Zhang G, Abbeel P, Levine S. MCP: Learning composable hierarchical control with multiplicative compositional policies. Adv Neural Inf Process Syst. 2019; 32.

392. Ren J, Li Y, Ding Z, Pan W, Dong H. Probabilistic mixture-of-experts for efficient deep reinforcement learning. arXiv. 2021; 2104.09122. https://doi.org/10.48550/arXiv.2104.09122

393. Yang Z, Ren K, Luo X, Liu M, Liu W, Bian J, Zhang W, Li D. Towards applicable reinforcement learning: improving the generalization and sample efficiency with policy ensemble. arXiv. 2022; 2205.09284. https://doi.org/10.48550/arxiv.2205.09284

394. Cheng G, Dong L, Cai W, Sun C. Multi-task reinforcement learning with attention-based mixture of experts. IEEE Robot Autom Lett. 2023; 8(6): 3812–3819. https://doi.org/10.1109/lra.2023.3271445

395. Hendawy A, Peters J, D'Eramo C. Multi-task reinforcement learning with mixture of orthogonal experts. arXiv. 2023; 2311.11385. https://doi.org/10.48550/arxiv.2311.11385

396. McIntosh TR, Susnjak T, Liu T, Watters P, Halgamuge MN. From Google Gemini to OpenAI Q* (Q-star): a survey of reshaping the generative artificial intelligence (AI) research landscape. arXiv. 2023; 2312.10868. https://doi.org/10.48550/arxiv.2312.10868

397. Brooks RA. New approaches to robotics. Science. 1991; 253(5025): 1227–1232. https://doi.org/10.1126/science.253.5025.1227 PMID: 17831441

398. Steels L, Brooks R. editors. The artificial life route to artificial intelligence: building embodied, situated agents. London, United Kingdom: Routledge; 1995. https://doi.org/10.4324/9781351001885

399. Pezzulo G, Barsalou LW, Cangelosi A, Fischer MH, McRae K, Spivey MJ. The mechanics of embodiment: a dialog on embodiment and computational modeling. Front Psychol. 2011; 2: 5. https://doi.org/10.3389/fpsyg.2011.00005 PMID: 21713184

400. Kober J, Bagnell JA, Peters J. Reinforcement learning in robotics: a survey. Int J Rob Res. 2013; 32 (11): 1238–1274. https://doi.org/10.1177/0278364913495721

401. Kormushev P, Calinon S, Caldwell DG. Reinforcement learning in robotics: applications and real-world challenges. Robotics. 2013; 2(3): 122–148. https://doi.org/10.3390/robotics2030122

402. Pezzulo G, Barsalou LW, Cangelosi A, Fischer MH, McRae K, Spivey MJ. Computational grounded cognition: a new alliance between grounded cognition and computational modeling. Front Psychol. 2013; 3: 612. https://doi.org/10.3389/fpsyg.2012.00612 PMID: 23346065

403. Lee SW, Seymour B. Decision-making in brains and robots—the case for an interdisciplinary approach. Curr Opin Behav Sci. 2019; 26: 137–145. https://doi.org/10.1016/j.cobeha.2018.12.012

404. Neftci EO, Averbeck BB. Reinforcement learning in artificial and biological systems. Nature Machine Intelligence. 2019; 1(3): 133–143. https://doi.org/10.1038/s42256-019-0025-4

405. Wilson M. Six views of embodied cognition. Psychon Bull Rev. 2002; 9(4): 625–636. https://doi.org/10.3758/bf03196322 PMID: 12613670

406. Barsalou LW. Grounded cognition. Annu Rev Psychol. 2008; 59: 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639 PMID: 17705682

407. Filliter JH, Glover JM, McMullen PA, Salmon JP, Johnson SA. The DalHouses: 100 new photographs of houses with ratings of typicality, familiarity, and degree of similarity to faces. Behav Res Methods. 2016; 48(1): 178–183. https://doi.org/10.3758/s13428-015-0561-8 PMID: 25675877

408. Witten IH. An adaptive optimal controller for discrete-time Markov environments. Inf Control. 1977; 34 (4): 286–295. https://doi.org/10.1016/s0019-9958(77)90354-0

409. Barto AG, Sutton RS, Anderson CW. Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans Syst Man Cybern. 1983; 13(5): 834–846. https://doi.org/10.1109/tsmc.1983.6313077

410. Sutton RS. Temporal credit assignment in reinforcement learning [Doctoral dissertation]. Amherst (MA): University of Massachusetts, Amherst; 1984.

411. Sutton RS. Learning to predict by the methods of temporal differences. Mach Learn. 1988; 3(1): 9–44. https://doi.org/10.1007/bf00115009

412. Dayan P. The convergence of TD(λ) for general λ. Mach Learn. 1992; 8(3–4): 341–362. https://doi.org/10.1023/a:1022632907294

413. Dayan P, Sejnowski TJ. TD(λ) converges with probability 1. Mach Learn. 1994; 14(3): 295–301. https://doi.org/10.1023/a:1022657612745

414. Watkins CJ. Learning from delayed rewards [Doctoral dissertation]. Cambridge, United Kingdom: University of Cambridge; 1989.

415. Watkins CJ, Dayan P. Q-learning. Mach Learn. 1992; 8(3–4): 279–292. https://doi.org/10.1007/bf00992698

416. Rummery GA, Niranjan M. On-line Q-learning using connectionist systems. Cambridge, United Kingdom: Department of Engineering, University of Cambridge; 1994. Technical Report No.: CUED/F-INFENG/TR 166.

417. Li J, Schiller D, Schoenbaum G, Phelps EA, Daw ND. Differential roles of human striatum and amygdala in associative learning. Nat Neurosci. 2011; 14(10): 1250–1252. https://doi.org/10.1038/nn.2904 PMID: 21909088

418. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. Econometrica. 1979; 47 (2): 263–291. https://doi.org/10.2307/1914185

419. Carandini M., & Heeger D. J. (2012). Normalization as a canonical neural computation. Nat Rev Neurosci. 2012; 13(1): 51–62. https://doi.org/10.1038/nrn3136

420. Rangel A, Clithero JA. Value normalization in decision making: theory and evidence. Curr Opin Neurobiol. 2012; 22(6): 970–981. https://doi.org/10.1016/j.conb.2012.07.011 PMID: 22939568

421. Palminteri S, Lebreton M. Context-dependent outcome encoding in human reinforcement learning. Curr Opin Behav Sci. 2021; 41: 144–151. https://doi.org/10.1016/j.cobeha.2021.06.006

422. Barraclough DJ, Conroy ML, Lee D. Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci. 2004; 7(4): 404–10. https://doi.org/10.1038/nn1209 PMID: 15004564

423. Morita K, Kato A. Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. Front Neural Circuits. 2014; 8: 36. https://doi.org/10.3389/fncir.2014.00036 PMID: 24782717

424. Kato A, Morita K. Forgetting in reinforcement learning links sustained dopamine signals to motivation. PLOS Comput Biol. 2016; 12(10): e1005145. https://doi.org/10.1371/journal.pcbi.1005145 PMID: 27736881

425. Katahira K, Yuki S, Okanoya K. Model-based estimation of subjective values using choice tasks with probabilistic feedback. J Math Psychol. 2017; 79: 29–43. https://doi.org/10.1016/j.jmp.2017.05.005

426. Toyama A, Katahira K, Ohira H. A simple computational algorithm of model-based choice preference. Cogn Affect Behav Neurosci. 2017; 17(4): 764–783. https://doi.org/10.3758/s13415-017-0511-2 PMID: 28573384

427. Klopf AH. Brain function and adaptive systems—a heterostatic theory. Bedford (MA): Air Force Cambridge Research Laboratories; 1972. Technical Report No.: AFCRL-72-0164.

428. Sutton RS, Barto AG. Toward a modern theory of adaptive networks: expectation and prediction. Psychol Rev. 1981; 88(2): 135–170. https://doi.org/10.1037/0033-295X.88.2.135 PMID: 7291377

429. Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika. 1933; 25(3/4): 285–294. https://doi.org/10.2307/2332286

430. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006; 441(7095): 876–879. https://doi.org/10.1038/nature04766

431. Cohen JD, McClure SM, Yu AJ. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philos Trans R Soc Lond B Biol Sci. 2007; 362(1481): 933–942. https://doi.org/10.1098/rstb.2007.2098 PMID: 17395573

432. Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD. Humans use directed and random exploration to solve the explore-exploit dilemma. J Exp Psychol Gen. 2014; 143(6): 2074–2081. https://doi.org/10.1037/a0038199 PMID: 25347535

433. Speekenbrink M, Konstantinidis E. Uncertainty and exploration in a restless bandit problem. Top Cogn Sci. 2015; 7(2): 351–367. https://doi.org/10.1111/tops.12145 PMID: 25899069

434. Gershman SJ. Deconstructing the human algorithms for exploration. Cognition. 2018; 173: 34–42. https://doi.org/10.1016/j.cognition.2017.12.014 PMID: 29289795

435. Schulz E, Gershman SJ. The algorithmic architecture of exploration in the human brain. Curr Opin Neurobiol. 2019; 55: 7–14. https://doi.org/10.1016/j.conb.2018.11.003 PMID: 30529148

436. Nelder JA, Mead R. A simplex method for function minimization. Comput J. 1965; 7(4): 308–313. https://doi.org/10.1093/comjnl/7.4.308