













RESEARCH ARTICLE

Reinforcement learning with associative or discriminative generalization across states and actions: fMRI at 3 T and 7 T

Jaron T. Colas^{1,2,3}  | Neil M. Dundon^{1,4}  | Raphael T. Gerraty^{5,6,7}  |
 Natalie M. Saragosa-Harris^{8,9}  | Karol P. Szymula¹⁰  | Koranis Tanwisuth^{2,11}  |
 J. Michael Tyszka²  | Camilla van Geen^{6,12}  | Harang Ju¹³  |
 Arthur W. Toga¹⁴  | Joshua I. Gold¹⁵  | Dani S. Bassett^{10,16,17,18,19,20}  |
 Catherine A. Hartley^{8,21}  | Daphna Shohamy^{5,6,22}  | Scott T. Grafton¹  |
 John P. O'Doherty^{2,3} 

¹Department of Psychological and Brain Sciences, University of California, Santa Barbara, California, USA

²Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California, USA

³Computation and Neural Systems Program, California Institute of Technology, Pasadena, California, USA

⁴Department of Child and Adolescent Psychiatry, Psychotherapy, and Psychosomatics, University of Freiburg, Freiburg im Breisgau, Germany

⁵Department of Psychology, Columbia University, New York, New York, USA

⁶Zuckerman Mind Brain Behavior Institute, Columbia University, New York, New York, USA

⁷Center for Science and Society, Columbia University, New York, New York, USA

⁸Department of Psychology, New York University, New York, New York, USA

⁹Department of Psychology, University of California, Los Angeles, California, USA

¹⁰Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA

¹¹Department of Psychology, University of California, Berkeley, California, USA

¹²Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Abstract

The model-free algorithms of “reinforcement learning” (RL) have gained clout across disciplines, but so too have model-based alternatives. The present study emphasizes other dimensions of this model space in consideration of associative or discriminative generalization across states and actions. This “generalized reinforcement learning” (GRL) model, a frugal extension of RL, parsimoniously retains the single reward-prediction error (RPE), but the scope of learning goes beyond the experienced state and action. Instead, the generalized RPE is efficiently relayed for bidirectional counterfactual updating of value estimates for other representations. Aided by structural information but as an implicit rather than explicit cognitive map, GRL provided the most precise account of human behavior and individual differences in a reversal-learning task with hierarchical structure that encouraged inverse generalization across both states and actions. Reflecting inference that could be true, false (i.e., overgeneralization), or absent (i.e., undergeneralization), state generalization distinguished those who learned well more so than action generalization. With high-resolution high-field fMRI targeting the dopaminergic midbrain, the GRL model's RPE signals (alongside value and decision signals) were localized within not only the striatum but also the substantia nigra and the ventral tegmental area, including specific effects of generalization that also extend to the hippocampus. Factoring in generalization as a multidimensional process in value-based learning, these findings shed light on complexities that, while challenging classic RL, can still be resolved within the bounds of its core computations.

Scott T. Grafton and John P. O'Doherty are co-senior authors.

[Correction added on 22 September 2022, after first online publication: Typographical errors were corrected in the article and Supporting Information was updated without changing substantial content].

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Human Brain Mapping* published by Wiley Periodicals LLC.

¹³Neuroscience Graduate Group, University of Pennsylvania, Philadelphia, Pennsylvania, USA

¹⁴Laboratory of Neuro Imaging, USC Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, Los Angeles, California, USA

¹⁵Department of Neuroscience, University of Pennsylvania, Philadelphia, Pennsylvania, USA

¹⁶Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA

¹⁷Department of Neurology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

¹⁸Department of Psychiatry, University of Pennsylvania, Philadelphia, Pennsylvania, USA

¹⁹Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania, USA

²⁰Santa Fe Institute, Santa Fe, New Mexico, USA

²¹Center for Neural Science, New York University, New York, New York, USA

²²Kavli Institute for Brain Science, Columbia University, New York, New York, USA

Correspondence

Jaron T. Colas, Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA, USA.
Email: jcolas@ucsb.edu

Funding information

Army Research Office, Grant/Award Numbers: W911NF-18-1-0244, W911NF-16-1-0474, W911NF-19-2-0026; Klingenstein-Simons Neuroscience Fellowship; National Institute for Mathematical and Biological Synthesis; National Institute of Biomedical Imaging and Bioengineering, Grant/Award Number: P41 EB015922; National Institute of Mental Health, Grant/Award Numbers: P50 MH094258, R01 MH115557; National Institute on Drug Abuse, Grant/Award Number: R01 DA040011

KEYWORDS

cognitive map, counterfactual learning, dopaminergic midbrain, generalization, hippocampus, individual differences, model-free and model-based, multifield fMRI, reinforcement learning, striatum

1 | INTRODUCTION

“Reinforcement learning” (RL) is a successful computational framework for describing the means by which an agent can learn from feedback in their environment to select actions that maximize future reward. This framework has been canonized not only in machine learning and artificial intelligence (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998) but also in psychology (Bush & Mosteller, 1951a; Rescorla & Wagner, 1972) and neuroscience (Montague et al., 1996; Schultz, 2015; Schultz et al., 1997). In computational modeling of the nervous system, the discovery that the phasic activity of dopamine neurons represents the signature reward-prediction error (RPE) has firmly placed RL at the core of our understanding of the neurobiological basis of reward-related learning. Yet, as canonical RL models of the model-free variety have proliferated to rise to the challenges of learning, so too have model-based

alternatives to the RL framework as well as dual-systems models that are both model-free and model-based (Daw et al., 2005; Doll et al., 2012; O'Doherty et al., 2017, 2021). It is compellingly intuitive to consider these counterparts in terms of a straightforward dichotomy: model-free versus model-based. Toward the simpler end of the spectrum, model-free processes entail implicit caching of learned associations; toward the more complex end of the spectrum, model-based processes instead construct explicit cognitive models of the environment (which are incidentally agnostic with respect to conscious awareness). On the other hand, the present study expands this model space by introducing two additional dichotomies in their own right: associative versus discriminative generalization and state versus action generalization.

In applying RL to neurobiology, previous approaches have typically treated states and actions in the world as independent events, such that knowledge acquired about one state or action does not

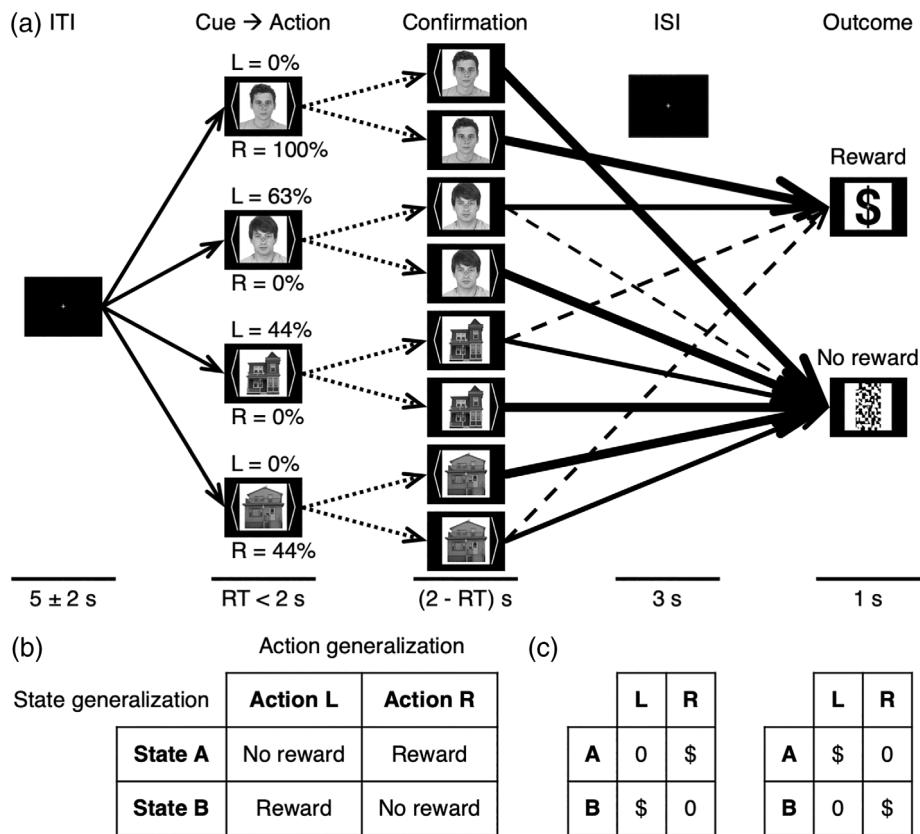


FIGURE 1 Task. (a) This schematic of the hierarchical reversal-learning task performed during fMRI scanning includes the probabilities of a rewarded outcome in one of 12 blocks. Following an intertrial interval (ITI) with a fixation cross, one of four paired states (i.e., cues) was presented with equal 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 model-based algorithm is an explicit model but for generalized reinforcement learning is an implicit model.

inform the agent about other states or actions. However, virtually all dynamic environments in the real world can be characterized by connections and patterns in relational structure across events that are disconnected not only temporally but also episodically, where “episodes” here correspond to perceived groupings within a sequence (and presently discrete trials within the experiment). As an agent grapples with uncertainty, such structured interdependence means that information obtained about one state or action can provide more general knowledge about other states and actions as well. A prototypical solution for leveraging the additional information is counterfactual inference. The credit assignment of standard RL models does not account for this interdependence or for mechanisms mediating the generalization it may support.

The present study operationalizes the concepts of associative versus discriminative generalization in relation to implicitly inferential

counterfactual learning (cf. Aquino et al., 2020; Balcarras & Womelsdorf, 2016; Ballard et al., 2019; Baram et al., 2021; Charpentier et al., 2020; Collette et al., 2017; Daw & Shohamy, 2008; Gläscher et al., 2009; Hampton et al., 2007; Hauser et al., 2014, 2015; Lesage & Verguts, 2021; Liu et al., 2021; Matsumoto et al., 2007; Mattar & Daw, 2018; Reiter et al., 2017; Vinckier et al., 2016; Wimmer et al., 2012; Zaki et al., 2016) that also differs with respect to states versus actions. The simpler associative generalization treats different representations as if they were equivalent or at least similar, which can but does not necessarily imply inference. In contrast, the more complex discriminative generalization treats different representations as if they were linked but specifically not equivalent—effectively implying a sort of emergent model (or cognitive map) with more abstract credit assignment. Being semi-inferential, this counterfactual learning entails value updating that occurs without direct

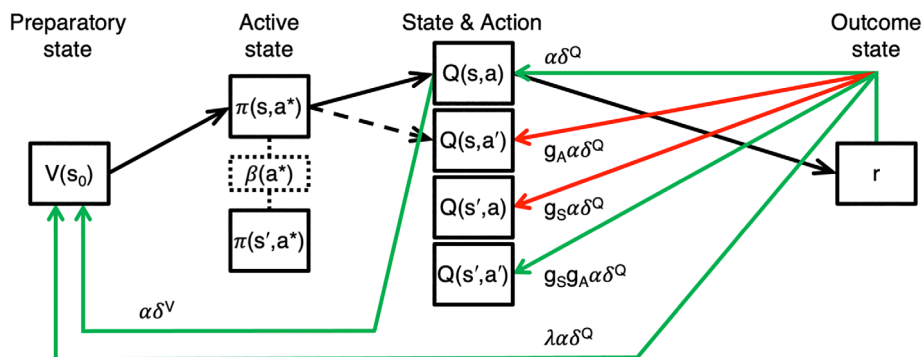


FIGURE 2 The “generalized reinforcement learning” (GRL) model. Compare to Figure 1. Here depicted in its 7-parameter form, the GRL model introduces the concepts of state generalization (g_S) and action generalization (g_A) as enhancements of the “critic/Q-learner” model, which is represented by the case where $g_A = g_S = 0$. The agent begins the trial in the preparatory state s_0 having a state value $V(s_0)$ cached by the “critic” module. At trial onset, the agent is presented with a random active state s having a complementary state s' within the same category (e.g., faces). The agent’s probabilistic action-selection policy $\pi(s, a^*)$ over available actions a^* is determined by not only their respective action values $Q(s, a^*)$, which are cached by the “Q-learner” module, but also action-specific bias and hysteresis $\beta(a^*)$. For this example, the agent’s chosen action a corresponds to the greater action value $Q(s, a)$ that updates $V(s_0)$ via a state-value-prediction error δ^V weighted by the learning rate α , which follows from this temporal-difference (TD) algorithm that also tracks passive states. The outcome of the action is a reward r that updates $Q(s, a)$ by way of an action-value-prediction error δ^Q . With the “TD(λ)” eligibility trace, $V(s_0)$ is also updated a second time by this same reward-prediction error (RPE) but reweighted with a decay multiplier as the eligibility parameter λ . With analogy to the temporal generalization of TD(λ), this new GRL model postulates that the generalized RPE signal is duplicated again, reweighted by g_A , and relayed to the action value $Q(s, a')$ representing the complementary nonchosen action a' within state s . Likewise, the generalized RPE is reweighted by g_S and relayed to the action value $Q(s', a)$ representing the chosen action a for complementary state s' . Finally, these parameters interact as a combined weight $g_S g_A$ modulating the RPE relayed to action value $Q(s', a')$ for the complementary action a' within the complementary state s' . Positive value updates are indicated with green arrows for this rewarded example trial, whereas negative updates for the complementary state and the complementary action (but not their combination) are indicated with red arrows. The signs of these updates reflect discriminative generalization ($-1 \leq g_S < 0$ and $-1 \leq g_A < 0$), which is optimal here because of the anticorrelational structure across states and actions.

observation and so requires an internal representation of generalizable structure in the environment. Yet the influence of such an internal representation does not necessarily require a (cognitive) model-based process in the stricter senses of the term; inference of that level of comprehensiveness can be achieved with alternative algorithms that are also investigated here. Rather, fundamentally model-free signals could more simply be modulated as needed and relayed via generalization.

The centerpiece of this work is a “generalized reinforcement learning” (GRL) model that addresses specific aspects of generalization by efficiently exploiting correlational structure between both states and actions. Augmenting the model-free scheme in this parsimonious way is another approach amid a zeitgeist inclined to more sophisticated alternatives to the basic RL framework such as proper model-based control. Whereas this model still employs the temporal-difference (TD) prediction method (Dayan, 1992; Dayan & Sejnowski, 1994; Sutton, 1988; Sutton & Barto, 1998), the algorithm is modified—not only updating the estimated value of an experienced state or action but also generalizing so as to flexibly transfer value information to other states and actions related to what was experienced. As opposed to separate counterfactual RPE signals, this singular generalized RPE signal functions as a heuristic for relaying the counterfactual information that can be derived directly from immediate experience. To implement the present optimization for both behavioral and neural modeling, we introduced the different types of generalization in parallel as enhancements of the “critic/Q-learner”

model (Colas et al., 2017) that we previously developed and validated as a bridge between the “actor/critic” model (Barto et al., 1983, 2021; Sutton, 1984; Witten, 1977) and the “Q-learning” model (Watkins, 1989; Watkins & Dayan, 1992).

Suitably for testing this GRL model, a hierarchical reversal-learning paradigm (Figure 1) allowed for associative generalization but favored discriminative generalization across both states and actions. The tightly controlled task accomplished this with high-order structure imposed to link available actions as well as subsumed states to discriminate between within stimulus categories. This embedded task structure provided participants with opportunities to recognize and exploit patterns across related events in separate trials so as to maximize reward. To facilitate action generalization, only one of two actions would be rewarded per state as a rule. Moreover, to facilitate state generalization, states (i.e., visually discriminable cues) were also paired within a category such that opposite actions were rewarded between the two states. The rule within each state category thus defined a hierarchical metastate with mapping that could reverse independently of that for the other category’s binary metastate. The optimal strategy in this setting is inverse generalization that effectively infers and leverages anticorrelational interdependencies both between complementary actions within each state and between complementary states within each category.

Yet the task is difficult in a probabilistic and changing environment producing noisy input, and fully recognizing interdependencies across trials becomes nontrivial as working memory is taxed. These cognitive demands may instead predispose an “uncertain learner to implicit

generalization. Although the quasi-model-based GRL model does not include an explicit representation of the linkage between states or actions, the optimality of discriminative generalization follows from the potential for the generalizing agent to incorporate more of the available information and update value representations four times as frequently in this case (Figure 2). As an anticipatory strategy, this reduces uncertainty in preparation for the next encounter within either a given state or its category.

There is considerable precedent for inquiry surrounding generalization and structure in learning (Bush & Mosteller, 1951b; Ghirlanda & Enquist, 2003; Harlow, 1949; Shepard, 1957, 1987; Tenenbaum & Griffiths, 2001; Tversky, 1977), and this is even the case for the specific domain of value-based learning (Ballard et al., 2019; Baram et al., 2021; Behrens et al., 2018; Bernardi et al., 2020; Bromberg-Martin et al., 2010; Daw & Shohamy, 2008; Doll et al., 2012; Doll, Duncan, et al., 2015; Doll, Shohamy, & Daw, 2015; Gerraty et al., 2014; Gershman & Niv, 2015; Hampton et al., 2006, 2007; Karagoz et al., 2022; Kool et al., 2016, 2017, 2018; Lehnert et al., 2020; Liu et al., 2021; Mattar & Daw, 2018; O'Doherty, 2012; Park et al., 2020; Prévost et al., 2013; Sadacca et al., 2016; Schulz et al., 2020; Watanabe & Hikosaka, 2005; Wimmer et al., 2012; Wimmer & Shohamy, 2012; Wunderlich et al., 2011). Yet typical approaches have emphasized strictly associative forms of generalization based on equivalence or similarity; in the present context, these are actually counterproductive as conflation—that is, overgeneralization. Uniquely for the present study, its explication extends to individual learners and how discriminative generalization can manifest (or not manifest) across representations of both states and actions. The GRL model is sensitive to not only discriminative generalization, which is presently optimal, but also associative overgeneralization or simple undergeneralization, thereby capturing possible variability in how humans might generalize with true or false beliefs or just fail to generalize altogether. With an aim for pragmatism, efficiency, and flexibility rather than pure optimality, these parameterized forms of associative or discriminative generalization and state or action generalization were framed to dovetail with classic RL and its operating constraints in the midst of stochasticity and parallel effects of action-specific bias and hysteresis.

While serving to demonstrate the robustness of the techniques, this multisite study also allowed for a more diverse sample as part of the emphasis on individual differences (cf. Colas et al., 2017; Schönberg et al., 2007). Human participants performed one of two versions of the structured learning task while their brains were scanned with functional magnetic-resonance imaging (fMRI). The first experiment was conducted at a now-standard field strength (3 T) across five separate laboratories, whereas the second was conducted at a high (or “ultra-high”) field strength (7 T) in parallel so as to elucidate subtle neural signatures of the GRL model with high fidelity, introducing commensurable state-of-the-art imaging for a paradigm that lacks precedent for a high-field or multifold fMRI study (cf. Beisteiner et al., 2011; Colzoli et al., 2021; Da Costa et al., 2015; de Hollander et al., 2017; Morris et al., 2019; Sengupta et al., 2018; Theysohn et al., 2013; Torrisi et al., 2018; Zaretskaya et al., 2020; but

see Fontanesi, Gluth, Rieskamp, et al., 2019). While we did examine cortical signals, we primarily focused on subcortical regions of the basal ganglia that have been implicated in RL with evidence from earlier studies. Here the GRL model again benefits from the anchor of RL insofar as prior literature from the classic RL perspective still provides a firm foundation for further constraining hypotheses about signals in the brain. Bolstered by the advantages of high-field fMRI (De Martino et al., 2018; Dumoulin et al., 2018; Torrisi et al., 2018; Uğurbil, 2018), our neuroimaging protocols were optimized for higher spatial resolution to pinpoint RL and GRL mechanisms in not only the striatum but also the dopaminergic midbrain. The technical challenges posed by measurements within elusive dopaminergic nuclei (Barry et al., 2013; de Hollander et al., 2015, 2017; Düzel et al., 2009, 2015) were addressed by adopting tailored measures for image preprocessing and denoising.

The first hypothesis was that the GRL model offers a superior account of motivated behavior and especially the distribution of performance at the level of individual participants. This quasi-model-based extension of model-free RL could even stand to outcompete more unambiguously model-based solutions, including delta learning with a state-prediction error (SPE) (cf. Gläscher et al., 2010; Lee et al., 2014)—or here a “metastate-prediction error” (MPE)—as well as more sophisticated Bayesian inference with a hidden Markov model (HMM) (Ghahramani, 2001; cf. Hampton et al., 2006; Prévost et al., 2013). Second, this model was hypothesized to successfully capture dynamics of neural activity (O'Doherty et al., 2007) associated with the computations characterizing RL as implemented within mesostriatal circuits (Chase et al., 2015; Colas et al., 2017; Garrison et al., 2013; O'Doherty et al., 2003; O'Doherty et al., 2004; Pauli et al., 2015; Schönberg et al., 2007). Third, predictions for value signals in both the ventral striatum and ventromedial prefrontal cortex (vmPFC) (Bartra et al., 2013; Behrens et al., 2008; Chase et al., 2015; Clithero & Rangel, 2014; Colas et al., 2017; Gläscher et al., 2009; Hare et al., 2008; Jocham et al., 2011; Kim et al., 2006) were also tested alongside RPE signals, which poses a challenge because these two types of signals as well as decision signals are all interconnected. Fourth, targeting predictions entirely specific to the GRL model, interaction effects in the basal ganglia as well as the hippocampus were expected to reflect the relaying of learning signals to representations of other states and actions; these interactions could be between RPE signaling and state generalization or between RPE signaling and action generalization. The hippocampal formation of the medial temporal lobe is a viable candidate for representing not only spatial topological maps (Moser et al., 2008; O'Keefe & Nadel, 1978) but also cognitive maps (Lewin, 1935, 1936; Tolman, 1948) such as in this more abstract space of states and actions (Ballard et al., 2019; Baram et al., 2021; Behrens et al., 2018; Bernardi et al., 2020; Cazé et al., 2018; Daw & Shohamy, 2008; Gerraty et al., 2014; Liu et al., 2019, 2021; Mattar & Daw, 2018; Momennejad et al., 2018; Park et al., 2020; Schuck & Niv, 2019; Wimmer et al., 2012; Wimmer & Shohamy, 2012).

The aims of the present study thus include first replicating and then building upon the established narrative of RL in the human brain,

encompassing a trichotomy of value, decision, and learning signals. With a parsimoniously optimized implementation of the algorithmic template of RL, this flexible scheme for associative or discriminative generalization across states and actions broadens this narrative for predictably structured environments. Furthermore, with that narrative there arises an opportunity to reflect on how this generalization paradigm—as distinguished from one defined by a multistep task (Bellman, 1957; Daw et al., 2005, 2011; Gläscher et al., 2010; Lee et al., 2014; Sutton & Barto, 1998), for example—can relate to model-free, model-based, or quasi-model-based aspects of structural learning.

2 | RESULTS

The first version of the structured learning task included fMRI at 3 T and faces or houses as stimuli (16 in total), whereas the second version included high-resolution fMRI at 7 T and colors or directions of motion as stimuli (4 in total). These different versions were acquired in parallel, and the advantages of the differences in stimuli between them were twofold. The prosaic advantage applies to the 7-T fMRI data, which are more susceptible to signal dropout: Owing to discrepancies between the magnetic properties of the cerebrum and the cerebellum and the properties of the interstitial space between them, there is a risk of dropout in the vicinity of the fusiform gyrus and (to a lesser extent) the parahippocampal gyrus—that is, the fusiform face area (FFA) (Kanwisher et al., 1997) and the parahippocampal place area (PPA) (Epstein & Kanwisher, 1998), which would relate to processing of face and house stimuli (i.e., states), respectively. The more substantial advantage is that replicating both behavioral and neural results between somewhat different experiments rather than strictly identical experiments can speak to the robustness or generality of a given effect.

2.1 | Participant groups

Within each data set (i.e., the 3-T Face/House (“3FH”) version or the 7-T Color/Motion (“7CM”) version), the first step of the analysis entailed dividing participants into three subgroups according to model-independent performance on the task (Schönberg et al., 2007) as well as the results of model fitting (Colas et al., 2017) (Table 1). Learning performance could thus be related to both behavioral and neural aspects of the modeling for this difficult task. A subset of participants was initially set aside as the “Good learner” (“G”) group (3FH: $n = 31/47$; 7CM: $n = 16/22$) if choice accuracy was significantly greater than the chance level of 50% for a given individual ($p < .05$). The remaining participants for whom the null hypothesis of chance accuracy could not be rejected at the individual level ($p > .05$) were further subdivided between the “Poor learner” (“P”) group (3FH: $n = 9/47$; 7CM: $n = 5/22$) and the “Nonlearner” (“N”) group (3FH: $n = 7/47$; 7CM: $n = 1/22$) according to whether or not an RL model could yield a significant improvement in goodness of fit relative to a nested hysteresis model without sensitivity to reward or its omission. Despite additional free parameters, the hysteresis model was justified statistically as a baseline model superior to the chance or intercept models (Tables S1–S15).

As part of the overarching computational framework—that is, not only RL per se but also the associated policy for action selection—reaction time (RT) was measured to implicitly relate dynamical models of decision making (Busemeyer & Townsend, 1993; Colas, 2017; Laming, 1968; Luce, 1986; Ratcliff, 1978; Usher & McClelland, 2001) to this context of active value-based learning (Ballard & McClure, 2019; Fontanesi, Gluth, Spektor, et al., 2019; Fontanesi, Palminteri, & Lebreton, 2019; Frank et al., 2015; Luzzardo et al., 2017; McDougale & Collins, 2021; Miletić et al., 2020, 2021; Millner et al., 2018; Pedersen et al., 2017; Pedersen & Frank, 2020; Ratcliff & Frank, 2012; Sewell et al., 2019; Sewell & Stallman, 2020; Shahar

TABLE 1 Participant groups

	3-T Face/House			7-T Color/Motion		
	Good learner	Poor learner	Nonlearner	Good learner	Poor learner	Nonlearner
<i>n</i>	31	9	7	16	5	1
Accuracy (%)	62.8 (5.4)	50.1 (3.1)	50.1 (3.0)	62.3 (5.6)	49.8 (4.9)	50.0
Reaction time (ms)	974 (129)	757 (107)	671 (104)	989 (87)	784 (163)	1043
Missed trials (%)	4.2 (6.8)	8.1 (10.5)	9.5 (9.8)	10.1 (9.1)	12.5 (12.4)	30.7
Age (y)	26.0 (4.9)	25.1 (5.0)	23.9 (5.2)	26.9 (4.2)	31.8 (9.8)	27
Male:Female (%)	54.8	55.6	71.4	43.8	100	0

Note: A subset of participants was initially set aside as the “Good learner” group if choice accuracy was significantly greater than chance at the individual level ($p < .05$). The remaining participants with chance accuracy ($p > .05$) were assigned to either the “Poor learner” group or the “Nonlearner” group according to whether or not a learning model could yield an improvement in fit relative to a hysteresis model without sensitivity to learnable outcomes. A speed-accuracy tradeoff was exhibited between groups with concomitant effects in reaction time such that the Good-learner group was the slowest to respond ($p < .05$). Standard deviations are listed in parentheses below corresponding means.

et al., 2019; Viejo et al., 2015). As more difficult decisions were hypothesized to be slower (see below), so too were decisions made more conscientiously by more attentive learners. Regarding the latter hypothesis, both data sets exhibited a speed-accuracy tradeoff (Garrett, 1922; Johnson, 1939) between groups with effects in RT such that the Good-learner group with high accuracy was also the slowest to respond (3FH-GP: $M = 217$ ms, $t_{38} = 4.60$, $p < 10^{-4}$; 3FH-GN: $M = 304$ ms, $t_{36} = 5.81$, $p < 10^{-6}$; 7CM-GP: $M = 205$ ms, $t_{19} = 3.72$, $p < 10^{-3}$). Likewise, the Poor-learner group was marginally slower than the Nonlearner group (3FH-PN: $M = 87$ ms, $t_{14} = 1.63$, $p = .063$).

2.2 | Model comparison

The present GRL model has seven free parameters: two for basic RL (α , τ), three for action-specific bias and hysteresis (β_0 , λ_β , β_R), and two

for generalization (g_A , g_S) (see Section 4). This and 16 other models were formally compared with inclusion of a full factorial design permuting the novel factors of action generalization and state generalization while controlling for outcome-independent effects of action-specific bias and hysteresis (Colas et al., 2017). The candidates included basic model-free RL (1 model), quasi-model-based GRL (10 models), the model-based SPE (with delta learning) (1 model), the model-based HMM (with Bayesian learning) (2 models), and dual-systems models that are both model-free and model-based (3 models) (Table 2). (Note that the “state” determining the SPE or the HMM’s hidden state is not the cue itself but rather the cue category’s metastate for generalizable structure represented with two possibilities shown in Figure 1c.)

In this context, a negative sign for the action-generalization weight ($-1 \leq g_A < 0$) represents correct recognition of the fixed complementarity between available actions, thereby speeding up learning. Likewise, a negative sign for the state-generalization weight

Model	MF vs. MB	df	RL	GRL			SPE	HMM		Dual
			α	g_A	g_S	g_{SA}	α_{SPE}	θ_0	θ_1	w_{MB}
A0 S0	MF	5	α	–	–	–	–	–	–	–
A– S0	MF (~MB)	5	α	–1	–	–	–	–	–	–
AX S0	MF (~MB)	6	α	g_A	–	–	–	–	–	–
A0 S+	MF	5	α	–	+1	–	–	–	–	–
A0 S–	MF (~MB)	5	α	–	–1	–	–	–	–	–
A0 SY	MF (~MB)	6	α	–	g_S	–	–	–	–	–
A– S+	MF (~MB)	5	α	–1	+1	–1	–	–	–	–
A– S–	MF (~MB)	5	α	–1	–1	–1	–	–	–	–
AW SW	MF (~MB)	6	α	g	g	g	–	–	–	–
AX SY	MF (~MB)	7	α	g_A	g_S	g_A	–	–	–	–
AX SY Z	MF (~MB)	8	α	g_A	g_S	g_{SA}	–	–	–	–
SPE	MB	5	–	–	–	–	α_{SPE}	–	–	–
SPE+RL	MF + MB	7	α	–	–	–	α_{SPE}	–	–	w_{MB}
HMM0	MB	5	–	–	–	–	–	θ_0	–	–
HMM	MB	6	–	–	–	–	–	θ_0	θ_1	–
HMM0+RL	MF + MB	7	α	–	–	–	–	θ_0	–	w_{MB}
HMM+RL	MF + MB	8	α	–	–	–	–	θ_0	θ_1	w_{MB}

TABLE 2 Model parameters

Note: All of the learning models are listed in ascending order of complexity both within and across classes: RL is the most simple and followed by GRL, the state-prediction error (SPE) (i.e., metastate-prediction error or MPE), the hidden Markov model (HMM), and lastly the most complex dual-systems models. The algorithms are described in terms of being model-free (“MF”), model-based (“MB”), or quasi-model-based (“~MB”). Model-free and (cognitive) model-based learning rates are listed as α and α_{SPE} for the RPE and the SPE, respectively. For RL and GRL, the labels “A0”, “A–”, and “AX” denote the absence of action generalization ($g_A = 0$), maximally optimal discriminative action generalization ($g_A = -1$), and free action generalization ($-1 \leq g_A \leq 0$), respectively. The labels “S0”, “S+”, “S–”, and “SY” denote the absence of state generalization ($g_S = 0$), maximally suboptimal associative state generalization ($g_S = 1$), maximally optimal discriminative state generalization ($g_S = -1$), and free state generalization ($-1 \leq g_S \leq 1$), respectively. GRL model “AW|SW” is limited to a single free parameter g shared between these two types of generalization ($g_S = g$, $g_A = \min\{0, g_S\}$). GRL model “AX|SY|Z” adds a free parameter g_{SA} for an interaction term $g_S g_{SA}$ (i.e., $g_{SA} \neq g_A$). The HMM0 variant shares a consistency parameter θ_0 with the full HMM but omits the reversal rate ($\theta_1 = 0$). Dual-systems models (“MF + MB”) include a weighting parameter w_{MB} for the model-based system. “df” stands for degrees of freedom.

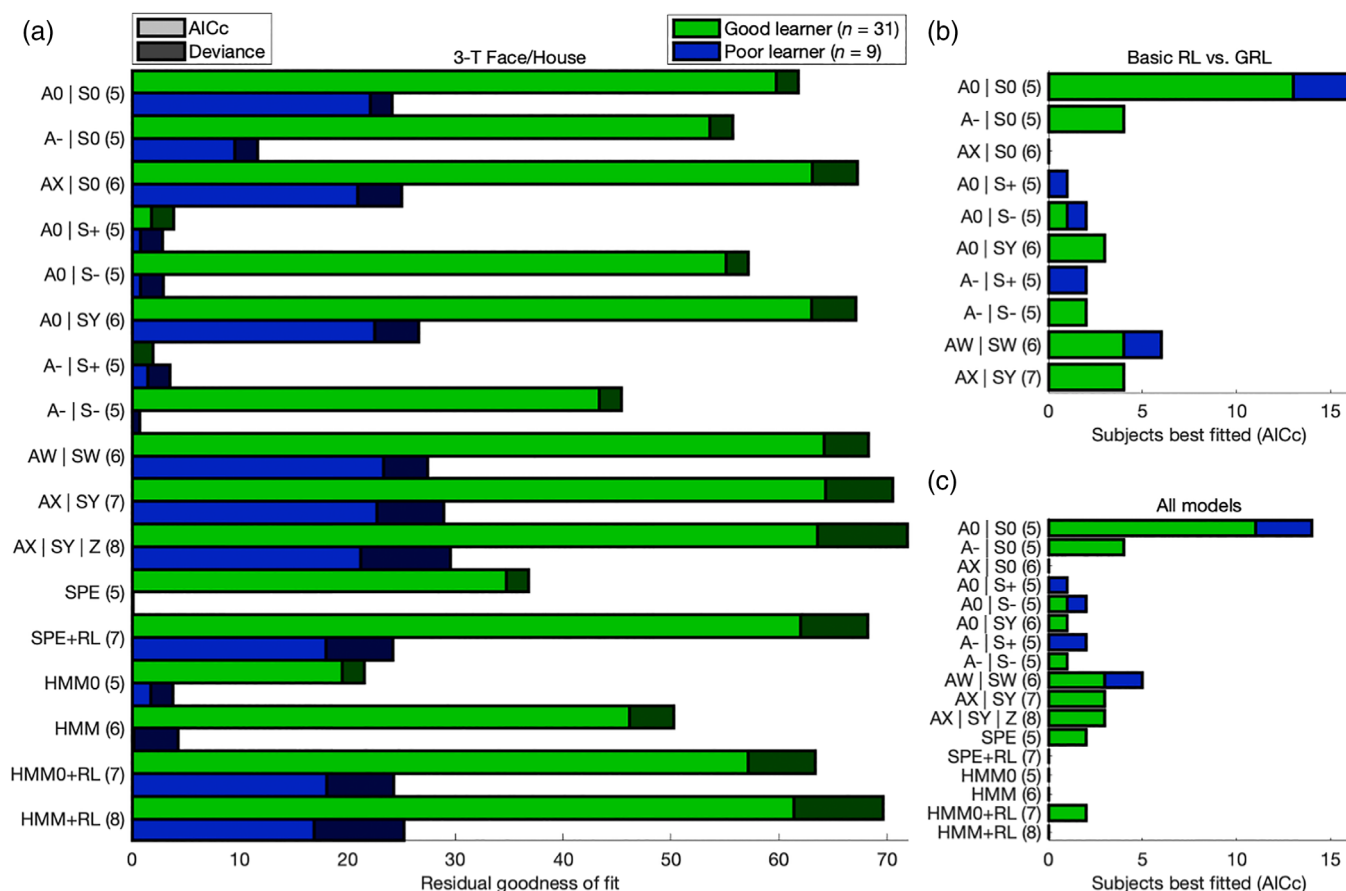


FIGURE 3 Model comparison: 3-T Face/House version. (a) For each learning model, average goodness of fit relative to the outcome-insensitive hysteresis model is shown with (light bars) and without (light and dark bars combined) a penalty for model complexity according to the corrected Akaike information criterion (AICc). The Good-learner (green bars) and Poor-learner (blue bars) groups are plotted separately. Emphasizing the result for the Good-learner group, the 7-parameter GRL model (“AX|SY”) outperformed all models even after correcting for model complexity and so justified inclusion of free parameters for both action and state generalization. The Nonlearner group is omitted here because these participants were best fitted by the hysteresis model following penalization. A more positive residual corresponds to a superior fit. Degrees of freedom are listed in parentheses. (b) Counts of the participants best fitted by the 7-parameter GRL model and each of its nested models according to the AICc are plotted with separation between learner groups, demonstrating that the majority strongly generalize and exhibit heterogeneity in generalization strategies within both groups. These trends could only be captured by a fully parameterized two-dimensional GRL model. (c) Broadening the scope to all models affirmed the preference for 7-parameter GRL and suggested negligible utilization of proper model-based strategies. This figure is related to Tables S1–S3.

$(-1 \leq g_s < 0)$ represents correct recognition of the fixed complementarity between paired states rewarding opposite actions within a category, whereas a positive sign $(0 < g_s \leq 1)$ instead represents incorrect overgeneralization across states within a category as if they were identical. The extremes for these parameters ($g_A = -1$, $g_S = 1$, or $g_S = -1$), their absence ($g_A = 0$ or $g_S = 0$), and their equivalence ($g_A = \min\{0, g_S\}$) were combined as alternatives to determine whether the two additional degrees of freedom were justified. Moreover, whereas the 7-parameter GRL model was expected to suffice with the assumption of an unparameterized interaction term $g_S g_A$, a version with an additional free parameter g_{SA} for an interaction term $g_S g_{SA}$ was also tested as part of due diligence.

Across the Good learners of both data sets, the 7-parameter GRL model including free parameters for both action and state generalization outperformed all nine models nested within it, the model that it

was nested within, and six model-based alternatives—even after correcting for model complexity according to the Akaike information criterion with correction for finite sample size (AICc) (Akaike, 1974; Hurvich & Tsai, 1989) (Figures 3a and 4a, Tables S1–S5). Crucially, when fitted at the level of individual subjects, this fully parameterized model could accommodate the heterogeneity in strategies for generalization observed within both Good-learner and Poor-learner groups—ranging from optimal (discriminative generalization) to semioptimal (undergeneralization) to suboptimal (associative overgeneralization) and with the majority strongly generalizing (3FH: $n = 24/40$; 7CM: $n = 17/21$) (Figures 3b/c and 4b/c). In other words, although a simpler alternative nested within the 7-parameter model may provide a decent account for some individuals, this more complex model in itself provided the most parsimonious account for the greatest proportion of participants.

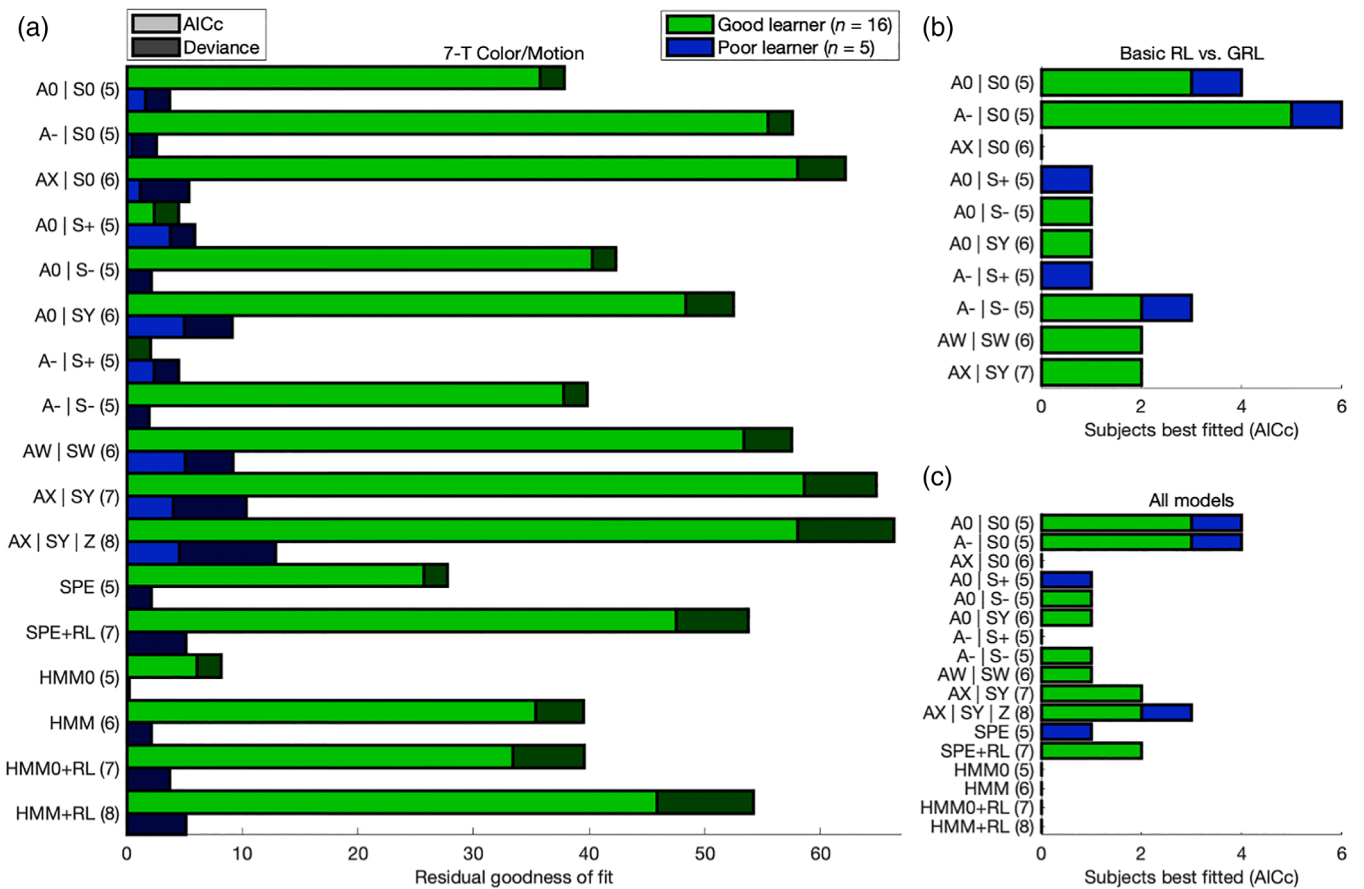


FIGURE 4 Model comparison: 7-T Color/Motion version. Compare to Figure 3. Results were replicated in the 7-T Color/Motion version of the experiment. This figure is related to Tables S4 and S5.

The lesser performance of the 8-parameter GRL model argues against an explanation reduced to mere overfitting because this further increase in complexity did not significantly improve performance. In an attempt to reduce complexity, a shared parameter for state and action generalization (i.e., $g_A = \min\{0, g_S\}$) proved insufficient as these orthogonal factors actually required separate free parameters. Notably for the approximation of a proper model-based strategy, the nested algorithm fixed with maximally optimal state and action generalization (i.e., $g_A = g_S = -1$) was not supported for most participants even when granted the benefits of fewer degrees of freedom. Likewise, the evidence did not favor the 5- or 6-parameter model-based algorithms (i.e., the SPE or the HMM) (cf. Aquino et al., 2020; Hampton et al., 2006; Prévost et al., 2013) or their respective dual-systems counterparts.

To again affirm the discriminability of the 7-parameter GRL model among both simpler and more complex alternatives, this entire pattern of results could be replicated after substituting simulated data generated by the fitted model itself (Figures S1 and S2, Tables S6–S10). Conversely, simulations generated with basic RL produced fitting results that instead aligned with basic RL (Figures S3 and S4, Tables S11–S15). That is, the complex model could be recovered from the complex model, and the simple model could be recovered from the simple model. This robust model discriminability rules out overfitting.

To complement the quantitative model comparison for overall goodness of fit, a posterior predictive check focused on a subset of diagnostic trials characterized by the purest effects of generalization. The hypothesis of generalized RL rather than basic RL could thus be tested at another level with qualitative falsification of the null hypothesis (Palminteri, Wyart, et al., 2017; Wilson & Collins, 2019). Based on parameter fits from the GRL model accommodating idiosyncratic generalization, Good and Poor learners were reclassified in “Discriminative generalizer” ($g_S < 0$) (3FH: $n = 19/40$; 7CM: $n = 12/21$), “Nongeneralizer” ($g_S = 0$) (3FH: $n = 11/40$; 7CM: $n = 2/21$), and “Associative generalizer” ($g_S > 0$) (3FH: $n = 10/40$; 7CM: $n = 7/21$) groups. The trials of interest corresponded to the first opportunities for generalization of reward within each block—that is, points in time before subsequent direct experience could update a value representation in the same direction as generalizable information would. After a given state-action pair was rewarded for the first time, the crucial test was whether the complementary action would correctly be chosen upon the next encounter with the complementary state within the same category. Despite an absence of direct reinforcement for the second state’s new reward contingencies—and even prior reinforcement to the contrary—the implicit inference of discriminative generalization nevertheless helps to boost this first-generalization accuracy above chance following indirect generalizable reinforcement.

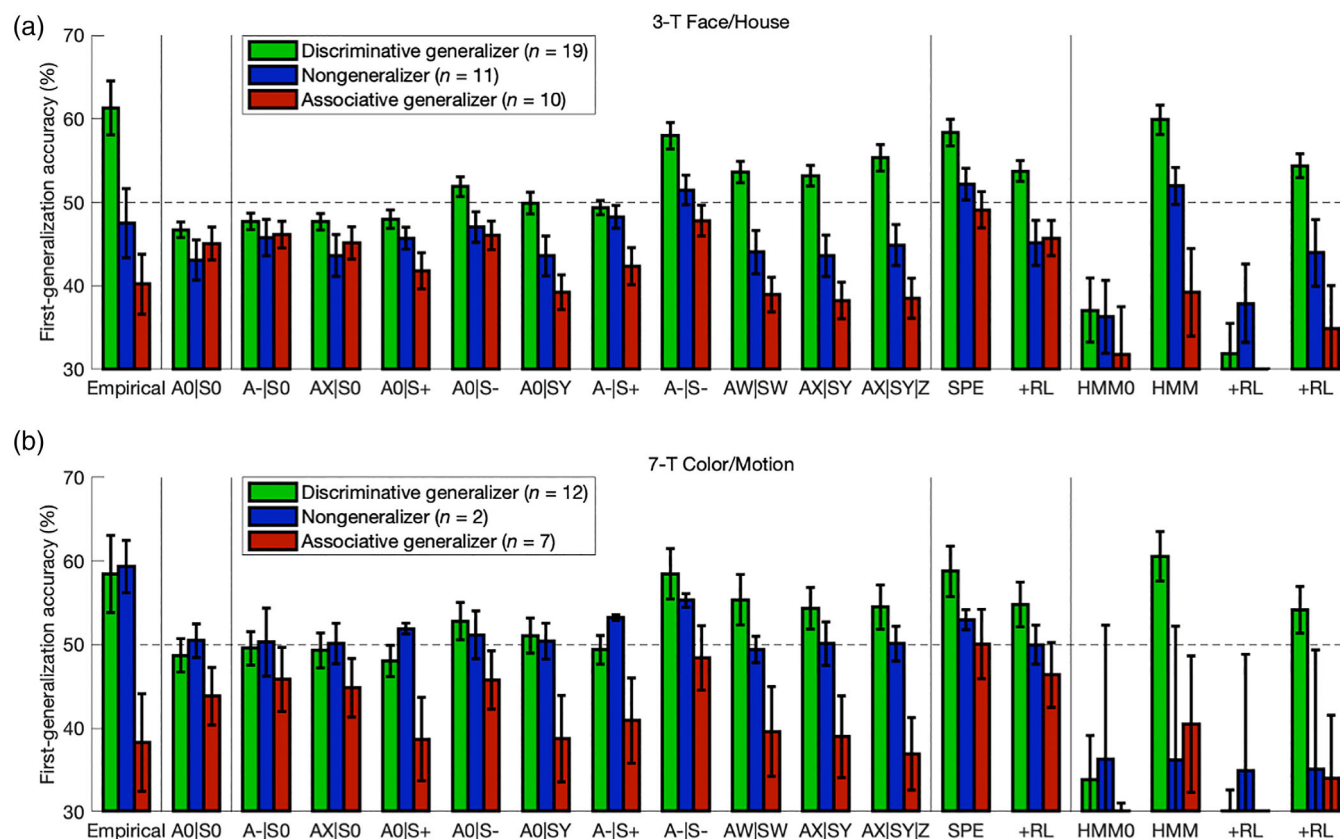


FIGURE 5 Posterior predictive check. (a) Focusing on only trials with the first opportunities for generalization of reward within each block, the purest effects of generalization were isolated in order to falsify basic RL. Using the GRL model, learners were reclassified in “Discriminative generalizer” ($g_s < 0$) (green bars), “Nongeneralizer” ($g_s = 0$) (blue bars), and “Associative generalizer” ($g_s > 0$) (red bars) groups. Only Discriminative generalizers could benefit from indirect reinforcement at these points in time, whereas Associative generalizers counterproductively overgeneralize the new information. In the absence of direct reinforcement for a complementary state’s newly correct action, Discriminative generalizers, Nongeneralizers, and Associative generalizers performed above ($p < .05$), at ($p > .05$), and below ($p < .05$) chance at first generalization, respectively. Simulated data sets from each competing model were yoked to their respective empirical data sets. Whereas the simpler nested models fail to account for this interaction effect between groups ($p > .05$), two-dimensional GRL models with state and action generalization capture the pattern successfully ($p < .05$). The (cognitive) model-based alternatives (SPE and HMM) and their respective dual-systems models (“+RL”) offered no substantial benefits here. (b) Results were replicated in the 7-T Color/Motion version of the experiment. Error bars indicate standard errors of the means.

Associative generalization counterproductively does the opposite in keeping with the false belief that the same action would be rewarded across the category.

As expected across both data sets, Discriminative generalizers did in fact perform above chance with the first generalization (3FH: $M = 11.3\%$, $t_{18} = 3.51$, $p = 10^{-3}$; 7CM: $M = 8.5\%$, $t_{11} = 1.84$, $p = .047$), whereas Associative generalizers performed not only below Discriminative generalizers (3FH: $M = 21.1\%$, $t_{27} = 4.09$, $p < 10^{-3}$; 7CM: $M = 20.1\%$, $t_{17} = 2.68$, $p = .008$) but also below chance (3FH: $M = 9.8\%$, $t_9 = 2.73$, $p = .012$; 7CM: $M = 11.7\%$, $t_6 = 2.00$, $p = .046$) (Figure 5). Moreover, the Nongeneralizer group’s first-generalization accuracy was not significantly below chance ($p > .05$) but was below that of Discriminative generalizers ($M = 13.8\%$, $t_{28} = 2.61$, $p = .007$). (Note that the subject counts of the generalization groups were not distributed as uniformly for the second sample, which left an insufficient Nongeneralizer group with a spuriously trending but nonsignificant result ($p > .05$) because of noise in the limited subset of trials that

were separated from the majority in this analysis.) Across all learners, first-generalization accuracy increased parametrically as the state-generalization weight was more negative (3FH: $r = 0.594$, $t_{38} = 4.55$, $p < 10^{-4}$; 7CM: $r = 0.459$, $t_{19} = 2.25$, $p = .018$). Altogether, hypotheses were confirmed across the board for these participant classifications derived from the GRL model.

Simulated data sets were generated with individually fitted instantiations of the computational models but yoked to their respective empirical data sets. That is, the simulated agents received input in silico according to what their respective participants actually encountered in the session. Regarding model comparison and falsification, this posterior predictive check confirmed that basic RL (sans generalization) was unable to account for the aforementioned effects, instead producing below-chance first-generalization accuracy across all groups of simulated agents ($p < .05$). Whereas the GRL agent has the capacity to infer this new category-level information prior to direct experience, the basic RL agent is limited to only information

TABLE 3 Parameters of the GRL model

	3-T Face/House			7-T Color/Motion		
	Good learner	Poor learner	Nonlearner	Good learner	Poor learner	Nonlearner
<i>n</i>	31	9	7	16	5	1
Accuracy (%)	62.8 (5.4)	50.1 (3.1)	50.1 (3.0)	62.3 (5.6)	49.8 (4.9)	50.0
Reward sensitivity $\log(\alpha(1-g_A-g_S+g_Sg_A)/\tau)$	0.058 (0.280)	-1.815 (2.309)	-1.823 (2.009)	0.069 (0.337)	-1.442 (2.266)	-5.764
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
Perseveration bias: Initial magnitude β_0	-0.066 (0.235)	-0.133 (0.438)	-0.169 (1.034)	-0.130 (0.153)	-0.393 (0.949)	-1.278
Alternation : Perseveration	21 : 10	4 : 5	4 : 3	13 : 3	3 : 2	1 : 0
Perseveration bias: Inverse decay rate λ_β	0.543 (0.371)	0.578 (0.404)	0.456 (0.421)	0.659 (0.318)	0.485 (0.403)	0.000
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
Intercept model: Residual deviance D_δ	78.56	48.67	16.75	73.97	42.15	22.28
Hysteresis model: Residual deviance D_δ	70.55	28.94	1.46	64.82	10.31	1.51

Note: Average fitted parameters for the preferred GRL model are listed for each participant group within each data set. Overall reward sensitivity was encapsulated by the ratio between generalized learning rates and temperature, which was greater for the Good-learner group than for the Poor-learner group in both data sets as expected ($p < .05$). Whereas action generalization did not differ between groups in either data set ($p > .05$), state generalization was more negative—that is, more optimal—for Good learners than for Poor learners ($p < .05$). The signs of individual fits are summarized 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; “alternation” ($\beta_0 < 0$) or “perseveration” ($\beta_0 > 0$) for hysteretic biases; and “leftward” or ($\beta_R < 0$) “rightward” ($\beta_R > 0$) for lateral biases. The residual deviance D_{df} (with degrees of freedom in the subscript) corresponds to the GRL model's improvement in fit relative to either null model. Standard deviations are listed in parentheses below corresponding means.

experienced within the current state that is at best neutral but may even be the opposite of what should be inferred.

Although simpler nested models with fixed generalization roughly approximate the empirical pattern in first-generalization accuracy, only the two-dimensional GRL models with parameterization of both state and action generalization could achieve the qualitative interaction effects within and between generalization-based groups ($p < .05$). (That the GRL model's fits to these trials are not quite perfect in terms of quantitative correspondence is merely a reflection of the fact that models were simultaneously fitted to the remaining 96% of trials along with the 4% emphasized at the moment; all of the trials were included in subsequent analyses that demonstrated the model's noteworthy quantitative precision.)

Remarkably, despite classification here being based on state generalization, the coexistence of action generalization was also essential for calibrating the model's recapitulated effects: Another example of subpar performance for the “AO|SY” model with $g_A = 0$ was evident in first-generalization accuracy remaining closer to chance among Discriminative generalizers ($p > .05$).

The more complex model-based algorithms (SPE and HMM) at best qualitatively matched GRL here but did not always perform as well—even as half of a dyad including basic RL. (The reduced HMMO variant in particular was limited by the rigidity of not explicitly representing reversals, such that new information that contradicts prior beliefs could not be integrated rapidly enough.) In this case, these alternatives did not offer any improvement that would justify

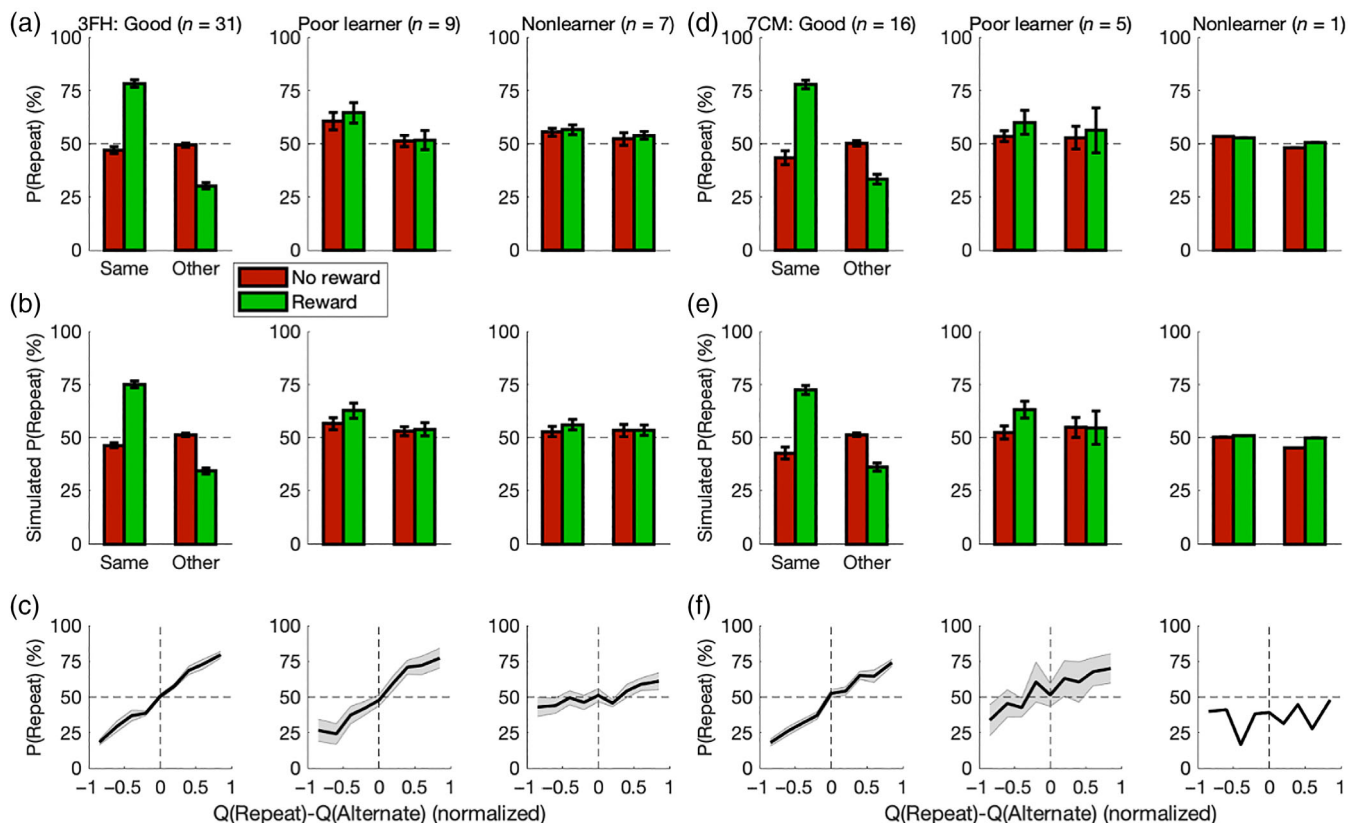


FIGURE 6 Behavioral modeling with the GRL model. (a) State generalization in particular relates to a qualitative pattern in choice behavior for this paradigm. Across all trials, conditions were defined by the most recent trials in which either the same (i.e., current) state was encountered or the other state within the current category was encountered; these trials were further binned according to whether the trial rewarded a given action (green bars) or provided no reward (red bars). While a rewarded action was more likely to be repeated for the same state among Good learners ($p < .05$), instead a nonrewarded action was more likely to be repeated from the other state ($p < .05$). This interaction effect ($p < .05$) follows from the complementarity of states within a category. Poor learners and Nonlearners did not exhibit such a pattern in behavior ($p > .05$). (b) To verify that the GRL model could reproduce these results with quantitative precision, simulated data sets were analyzed in the same fashion. (c) For all participant groups—including even “Nonlearners”—the probability of repeating the most recent action (independent of state) increased as a function of the difference between action values $Q_t(s_t, a)$ derived from the GRL model ($p < .05$). (d–f) Results were replicated in the 7-T Color/Motion version of the experiment. Error bars indicate standard errors of the means.

sacrificing the parsimony of GRL. All things being equal, Occam's razor would bias model selection away from the computational complexity demanded by a more model-based architecture as compared to a quasi-model-based but primarily model-free architecture that boasts simplicity. Unlike the addition of free parameters, this lack of parsimony—including a less straightforward neural implementation—is not readily quantifiable for formal penalization in proportion to the concomitant increase in model complexity.

2.3 | Behavioral modeling

With the model comparison pointing to the 7-parameter GRL model, the next steps were to further verify and interpret the individually fitted parameters of this model with reference to learning performance (Table 3). The model could first quantify overall reward sensitivity with a logarithmic transformation of the ratio between the sum of all four generalized learning rates and the softmax temperature

(cf. Colas et al., 2017; Schönberg et al., 2007). This sensitivity metric $\log(\alpha(1-g_A-g_S+g_Sg_A)/\tau)$ was greater for Good-learner groups than for Poor-learner groups across data sets (3FH: $M = 1.873$, $t_{38} = 4.54$, $p < 10^{-4}$; 7CM: $M = 1.510$, $t_{19} = 2.72$, $p = .007$). Likewise, choice accuracy increased parametrically with sensitivity across both learner groups (3FH: $r = 0.547$, $t_{38} = 4.03$, $p = 10^{-4}$; 7CM: $r = 0.645$, $t_{19} = 3.68$, $p < 10^{-3}$). In keeping with the speed-accuracy tradeoff, RT was analogously slower as sensitivity increased (with marginal significance for the latter data set) (3FH: $r = 0.506$, $t_{38} = 3.62$, $p < 10^{-3}$; 7CM: $r = 0.346$, $t_{19} = 1.61$, $p = .062$).

State generalization g_S was more negative—that is, more optimal—for Good learners than for Poor learners across data sets (3FH: $M = 0.551$, $t_{38} = 3.71$, $p < 10^{-3}$; 7CM: $M = 0.496$, $t_{19} = 1.89$, $p = .037$). Likewise, across all learners, choice accuracy increased as state generalization was more negative (3FH: $r = 0.509$, $t_{38} = 3.65$, $p < 10^{-3}$; 7CM: $r = 0.571$, $t_{19} = 3.04$, $p = .003$). Action generalization g_A did not differ between groups in either data set ($p > .05$). In other words, the Poor learners were primarily limited by difficulties with

properly discriminating and generalizing between states within a category rather than actions within a state—the former being the more complex process here. The dissociation between state generalization and action generalization was confirmed by the complete absence of any correlation between these parameters across all learners (3FH: $r = 0.006, p > .05$; 7CM: $r = -0.033, p > .05$).

These forms of discriminative generalization speed up learning across trials, and as alluded to previously, state generalization here relates to a qualitative pattern in choice behavior based on an interaction effect of reinforcement between hierarchically paired states (Figure 6a/d). Whereas the previous analysis was concerned with isolating generalization effects in a subset of trials, this analysis across all trials addressed a mixture of effects such as pure RL, generalized RL,

action-specific biases, and stochasticity from noise and exploration. Conditions for preceding outcomes were defined by the most recent trials in which either the same (i.e., current) state was encountered or the other, complementary state within the current category was encountered; these trials were further binned according to whether the trial rewarded a given action or provided no such reward. Among Good learners of either data set, a rewarded action was more likely to be repeated within the same state (3FH: $M = 31.3\%$, $t_{30} = 13.52$, $p = 10^{-14}$; 7CM: $M = 34.4\%$, $t_{15} = 8.34$, $p < 10^{-6}$); in contrast, a non-rewarded action was more likely to be repeated after being performed in the other state (3FH: $M = 19.2\%$, $t_{30} = 15.95$, $p < 10^{-15}$; 7CM: $M = 16.9\%$, $t_{15} = 9.08$, $p < 10^{-7}$), producing an interaction effect (3FH: $M = 50.5\%$, $t_{30} = 16.47$, $p < 10^{-16}$; 7CM: $M = 51.3\%$,

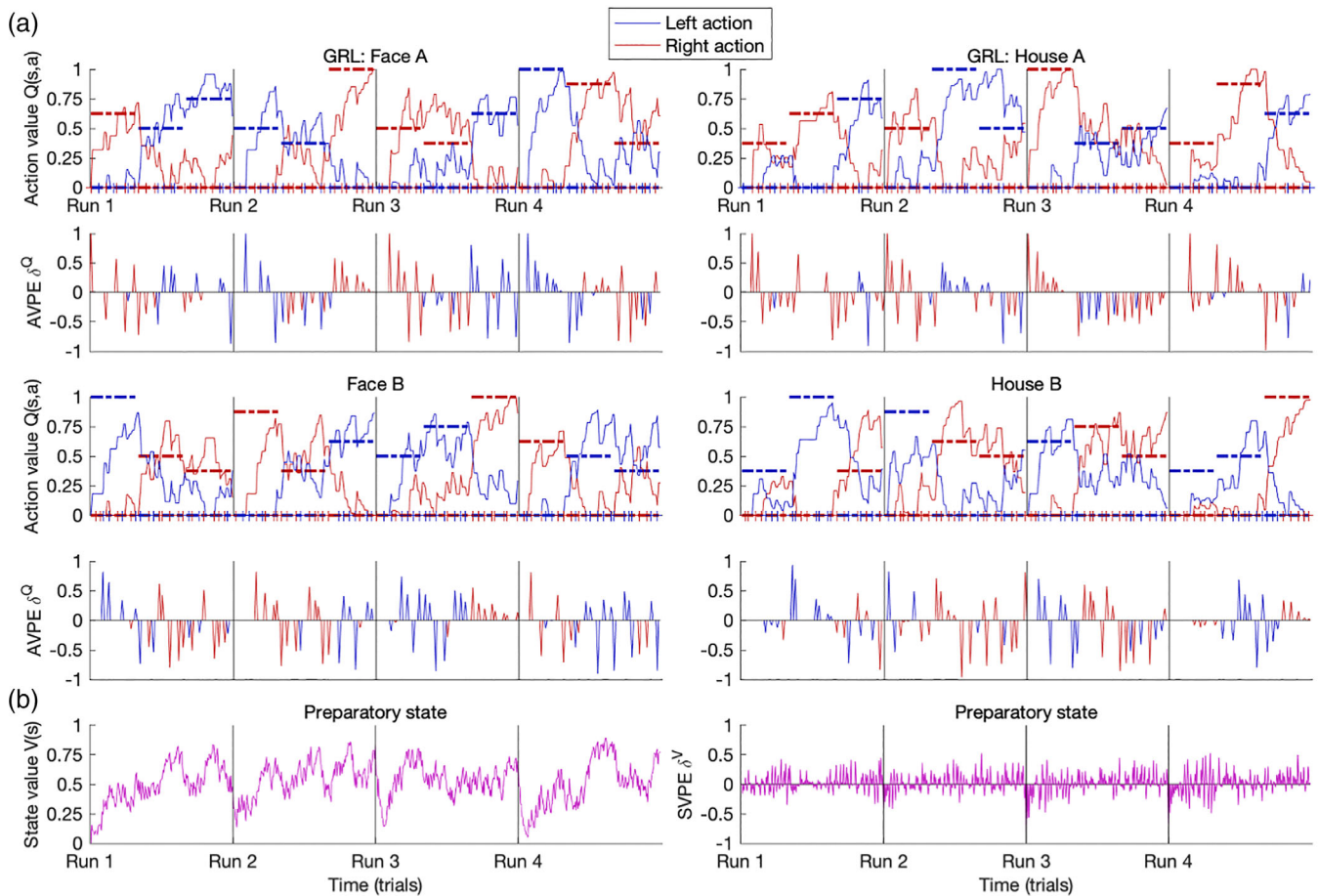


FIGURE 7 Predictions of the GRL model. Representative dynamics of value signals and learning signals generated by the GRL model are shown for the final participant in the Good-learner group of the 3-T Face/House data set. This modeling provided explicit quantitative predictions for internal decision variables within the (computational) model-based fMRI analysis. Parameters were assigned as follows for this participant: $\alpha = 0.318$, $g_A = -0.710$, $g_S = -0.808$, $\lambda = 0.500$, $\tau = 0.408$, $\beta_O = -0.067$, $\lambda_{\beta} = 0.753$, and $\beta_R = 0.178$. (a) 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 alongside 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_t^Q signals. Discriminative state and action generalization are evident with counterfactual updates of values for the three nonexperienced state-action pairs within a category (Figure S5). 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. (b) Whereas active states were tracked by the Q-learning component of this “critic/Q-learner” (CQ) model, the preparatory state preceding each active state was tracked by the CQ model's critic module for passive states. Essentially tracking the probability of reward for the entire task, the model's estimates of state values $V_t(s_0)$ are plotted alongside state-value-prediction error (SVPE) δ_t^V signals. From the temporal generalization of TD(λ), the value of the preparatory state was updated not only at the beginning of the trial but also at the end by way of the AVPE signal's eligibility trace.

$t_{15} = 9.46, p < 10^{-7}$). The Poor-learner and Nonlearner groups did not exhibit this pattern in their choices ($p > .05$). Identical analyses of simulations in a second posterior predictive check—the first being qualitative—confirmed that the GRL model could reproduce these results with quantitative precision (3FH-G: $p < .05$; 3FH-P: $p > .05$; 3FH-N: $p > .05$; 7CM-G: $p < .05$; 7CM-P: $p > .05$) (Figure 6b/e). By incorporating action-specific bias and hysteresis (Colas et al., 2017; Lau & Glimcher, 2005; Schönberg et al., 2007), this extended model simultaneously matched reward-independent effects on the dynamic base rates of action repetition or alternation as well.

Additional validation of model fitting could be found in (computational) model-based psychometric functions of choices and RTs; the former maps onto the standard softmax function embedded within the present model (Luce, 1959; Shepard, 1957; Sutton & Barto, 1998), and the latter has been shown to be generally applicable to RL (Ballard & McClure, 2019; Fontanesi, Gluth, Spektor, et al., 2019; Fontanesi, Palminteri, & Lebreton, 2019; Frank et al., 2015; Luzzardo et al., 2017; McDougle & Collins, 2021; Miletić et al., 2020, 2021; Millner et al., 2018; Pedersen et al., 2017; Pedersen & Frank, 2020; Ratcliff & Frank, 2012; Sewell et al., 2019; Sewell & Stallman, 2020; Shahar et al., 2019; Viejo et al., 2015). (Given the two-alternative forced choice, this logistic softmax model is nested within not only signal-detection theory (Green & Swets, 1966) but also the drift-diffusion model encompassing RT (Laming, 1968; Ratcliff, 1978; Ratcliff et al., 2016; Stone, 1960).) For all five participant groups across data sets—including even “Nonlearners” who actually do exhibit subtle signatures of learning—the probability of repeating the most recent action (independent of state) increased as a

function of the difference between action values $Q_t(s_t, a)$ derived from the GRL model (3FH-G: $\beta = 1.902, t_{30} = 9.66, p < 10^{-10}$; 3FH-P: $\beta = 1.986, t_8 = 2.70, p = .014$; 3FH-N: $\beta = 0.332, t_6 = 4.53, p = .002$; 7CM-G: $\beta = 1.668, t_{15} = 7.44, p = 10^{-6}$; 7CM-P: $\beta = 1.034, t_4 = 2.50, p = .033$) (Figure 6c/f). Along with effects on choices, RT became faster as the absolute difference between action values increased for 4 out of 5 participant groups (3FH-G: $\beta = 62 \text{ ms}, t_{30} = 2.78, p = .005$; 3FH-P: $\beta = 120 \text{ ms}, t_8 = 3.01, p = .008$; 3FH-N: $\beta = 107 \text{ ms}, t_6 = 3.07, p = .011$; 7CM-G: $p > .05$; 7CM-P: $\beta = 58 \text{ ms}, t_4 = 2.28, p = .042$). (The RT results for the 7-T Color/Motion version—one of which is the null result—are given less weight in consideration of the dynamic stimuli that require more time to recognize via perceptual decision making.)

2.4 | Neural substrates of the RL framework

Having demonstrated the efficacy of the GRL model and its fitted parameters with respect to behavior, a (computational) model-based analysis followed suit for the neuroimaging data (O'Doherty et al., 2007). For each participant and their experienced sequence of events, this modeling generated explicit quantitative predictions for internal decision variables (Figures 7 and S5). The tripartite neural model was characterized by (1) learning signals as the generalized RPEs from the GRL model, (2) value signals from the GRL model, and (3) decision-making signals as approximated by RT. Along with the hippocampus, the hypothesis space was constrained by focal regions of interest (ROIs) based on established

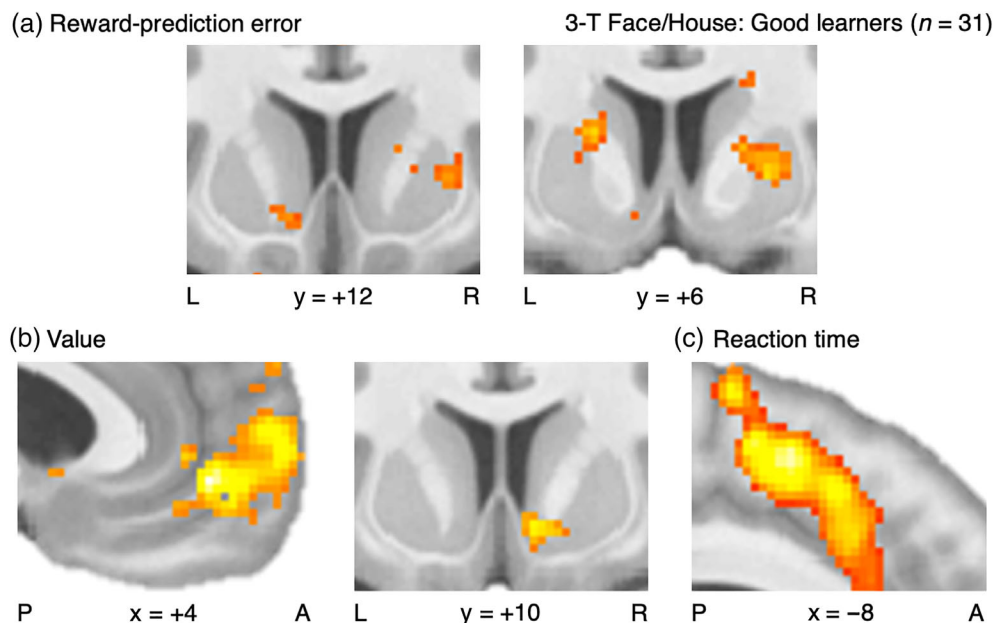


FIGURE 8 Neural substrates of the RL framework: 3-T Face/House version. (a) At 3 T, reward-prediction error (RPE) signals from the GRL model were significant at the set level (SVC $p_{FWE} < 0.05$) and identified throughout the striatum ($p < .005$), including the nucleus accumbens, the dorsal caudate nucleus, and the dorsal putamen (SVC $p_{FWE} < 0.05$). (b) Value signals from the GRL model were also significant at the set level (SVC $p_{FWE} < 0.05$) and identified in ventromedial prefrontal cortex (vmPFC) and the nucleus accumbens ($p < .005$, SVC $p_{FWE} < 0.05$). (c) As a proxy for decision-making signals, reaction time (RT) was associated with greater activity in medial frontal cortex (MFC) ($p < .005$, SVC $p_{FWE} < 0.05$). “L”, “R”, “P”, and “A” orient the left, right, posterior, and anterior directions, respectively. This figure is related to Tables S16 and S18.

precedents for the precise neural correlates of the RPE (Colas et al., 2017), subjective value (Bartra et al., 2013; Clithero & Rangel, 2014), and RT (Yarkoni et al., 2009). To further assess these neurophysiological signals in relation to learning performance evident in behavior, the participant groups were analyzed both collectively and separately for juxtaposition.

Reaction time served as a model-independent proxy for neural decision-making signals (Cisek, 2012; Cisek & Kalaska, 2010; Gold & Shadlen, 2007) that, with integration of sequential sampling, are characteristically ramping, bounded, and nonlinear (Colas, 2017; Usher & McClelland, 2001; Wang, 2002; Wong & Wang, 2006). Admittedly, limited temporal resolution translates to a risk of false positives at the level of interpretation when attempting to isolate decision signals among myriad other signals in the brain. Yet, although a measure of “time on task” is potentially relatable to constructs such as attention, arousal, difficulty, effort, engagement, or control, a longer RT essentially corresponds to greater cumulative neural activity for a dynamical decision-making process that is integrated across time (Carp et al., 2010; Colas, 2017; Grinband et al., 2011; Hare et al., 2011; Shenhav et al., 2014; Weissman & Carp, 2013; Yarkoni et al., 2009). Trial-by-trial RT is a more direct proxy for decision signals than a model-derived metric for normative difficulty such as the value difference—whether represented as the absolute difference $|Q(s,a_1) - Q(s,a_2)|$ (unsigned) or as chosen value minus nonchosen value $Q(s,a) - Q(s,a')$ (signed) (Colas, 2017).

For the 3-T Face/House images to first validate and expand the framework that the GRL model builds upon, analyses of the three key signals focused on the Good-learner group in consideration of their more robust task-relevant neural activity (Colas et al., 2017;

Schönberg et al., 2007) (Figure 8, Tables S16 and S18; see S18 for summary). That is, learning signals in the brain are clearest among those who consistently learn well as reflected in their behavior. Sets of ROIs were specified a priori for mesostriatal RPE signals (7 ROIs) and corticostriatal value signals (4 ROIs). The networks identified as encoding RPE or value signals were both significant at the set level for these ROIs (SVC $p_{FWE} < .05$). RPE signals from the GRL model were identified throughout the striatum ($p < .005$), including the nucleus accumbens, the dorsal caudate nucleus, and the dorsal putamen (SVC $p_{FWE} < .05$) (Figure 8a). Regarding the dopaminergic mid-brain, RPE signals were also observed in the substantia nigra (SN) ($p < .005$). Value signals from the GRL model were identified in vmPFC, the nucleus accumbens, and posterior cingulate cortex (PCC) ($p < .005$, SVC $p_{FWE} < 0.05$) (Figure 8b). In keeping with the decoupling of RPE and value signals in this paradigm, there were no common clusters in the striatum when testing for intersection of RPE and value networks ($p > .005$). Moreover, reaction time was associated with greater activity in medial frontal cortex (MFC) ($p < .005$, SVC $p_{FWE} < 0.05$) (Figure 8c).

The next portion of the fMRI analysis boasted greater spatial precision with high-resolution imaging for the 7-T Color/Motion data (Figures 9 and S6, Tables S17 and S18). The networks identified as encoding RPE or value signals were again both significant at the set level (SVC $p_{FWE} < 0.05$). RPE signals from the GRL model were localized within the SN and throughout the striatum ($p < .005$), including the nucleus accumbens (SVC $p_{FWE} < 0.05$) (Figures 9a and S6). Value signals from the GRL model were likewise identified in vmPFC and the nucleus accumbens ($p < .005$, SVC $p_{FWE} < 0.05$) (Figure 9b). Striatal RPE and value signals did not overlap here either ($p > .005$). For

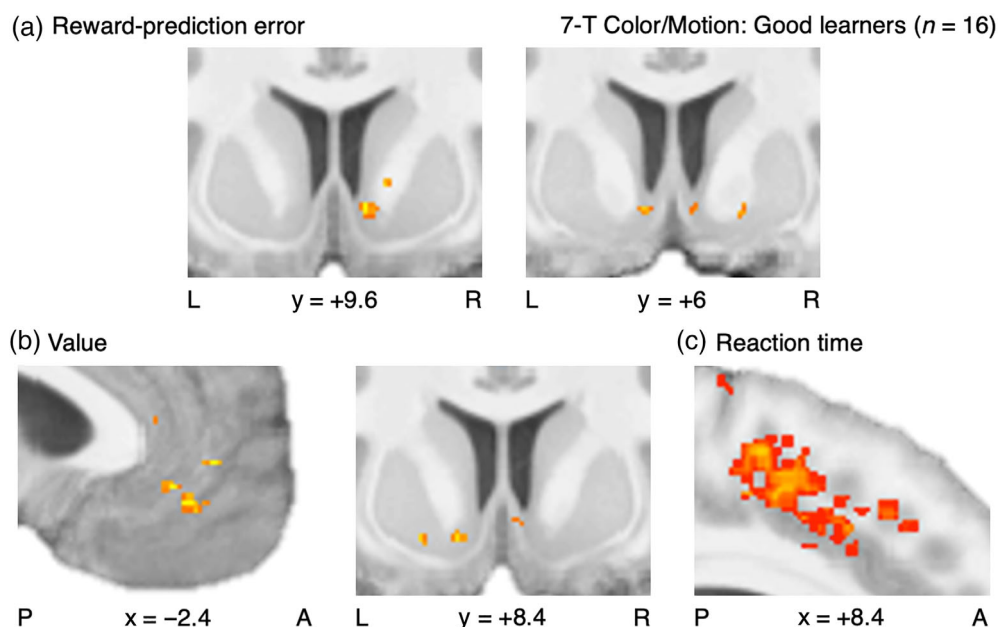


FIGURE 9 Neural substrates of the RL framework: 7-T Color/Motion version. (a) At 7 T, RPE signals from the GRL model were again significant at the set level (SVC $p_{FWE} < 0.05$) and identified throughout the striatum ($p < .005$), including the nucleus accumbens (SVC $p_{FWE} < 0.05$). (b) Value signals from the GRL model were again significant at the set level (SVC $p_{FWE} < 0.05$) and identified in vmPFC and the nucleus accumbens ($p < .005$, SVC $p_{FWE} < 0.05$). (c) RT was again associated with greater activity in MFC ($p < .005$, SVC $p_{FWE} < 0.05$). “L”, “R”, “P”, and “A” orient the left, right, posterior, and anterior directions, respectively. This figure is related to Figure S6 and Tables S17 and S18.

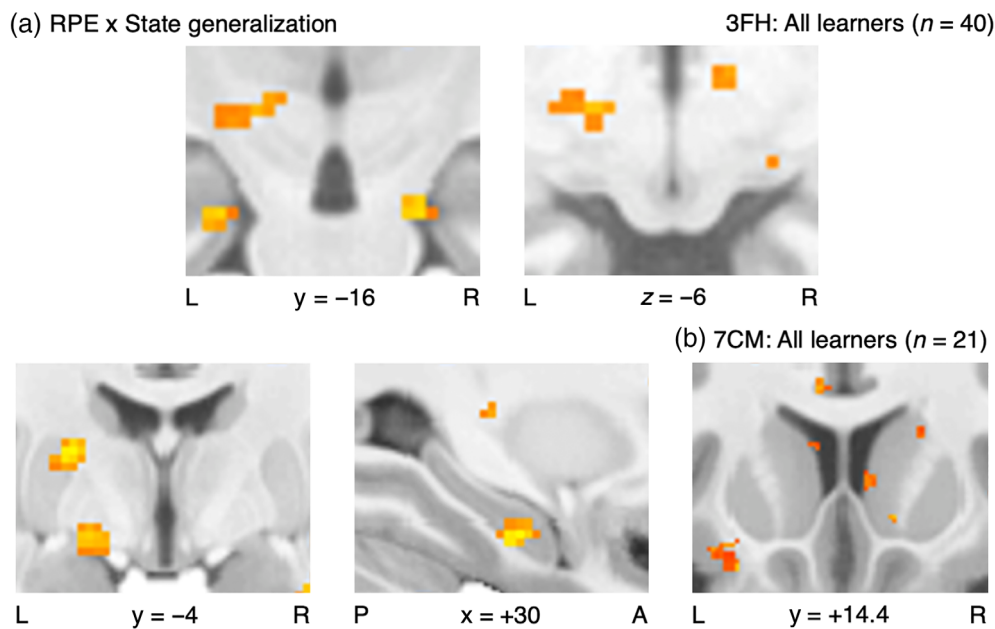


FIGURE 10 Neural substrates of the GRL model. (a) At 3 T, interaction effects between RPE signals and state generalization were significant at the set level (SVC $p_{FWE} < 0.05$) and identified in both the substantia nigra and the striatum ($p < .005$), including the posterior putamen in the vicinity of the dorsal caudate nucleus (SVC $p_{FWE} < 0.05$). In addition to mesostriatal circuits, generalization effects also modulated activity in the hippocampus ($p < .005$, SVC $p_{FWE} < 0.05$). (b) At 7 T, effects of state generalization were marginally significant at the set level (SVC $p_{FWE} < 0.10$) and identified throughout the striatum ($p < .005$), including the nucleus accumbens (SVC $p_{FWE} < 0.05$) and (with marginal significance) the anterior caudate nucleus (SVC $p_{FWE} < 0.10$). “L”, “R”, “P”, and “A” orient the left, right, posterior, and anterior directions, respectively. This figure is related to Figure S7 and Tables S19–S22.

yet another replication, RT was again associated with greater activity in MFC ($p < .005$, SVC $p_{FWE} < 0.05$) (Figure 9c).

2.5 | Neural substrates of the GRL model

Having elaborated on the RL framework within this paradigm, the second half of the neuroimaging analyses aimed to test additional predictions specific to the GRL model and as such entirely beyond the scope of basic RL. More precisely, interactions were tested for between RPE signals and either state generalization (i.e., $-g_S/\tau$) or action generalization (i.e., $-g_A/\tau$); such effects would be concordant with the hypothesis that there are relayed RPE signals mediating generalized updates of value representations that ultimately must interface with representations of states and actions such as in visual cortex and motor cortex, respectively (Lim et al., 2013; Magrabi et al., 2021; Philiastides et al., 2010). The topology of these representations and the relations between them is hypothesized to be encoded by a cognitive map maintained in the hippocampus, which would reflect downstream effects of generalized RPE signals from mesostriatal circuits without necessarily computing the RPE per se (cf. Ballard et al., 2019; Baram et al., 2021; Wimmer et al., 2012). These interaction effects were modeled with GRL parameters fitted at the level of individual subjects, including the temperature τ to factor in overall noise that diminishes the precision of the point estimates generated with the model's dynamics. Notably, the parameter for state generalization suggested greater emphasis given its greater intersubject variability and a more

direct link to successful learning, but the less variable factor of action generalization was also investigated.

The fundamental conceptual dissociation between states and actions (Averbeck & O'Doherty, 2022; Colas et al., 2017; O'Doherty et al., 2004) suggested an a-priori hypothesis that some mesostriatal and hippocampal circuitry would be uniquely implicated in either form of generalization. Accordingly, different categories of stimuli and different actions with different effectors evoked distinct neural representations that were amenable to fMRI by design—for example, engaging the FFA or the PPA with faces or houses, respectively. With implications for separable circuits for generalization, state and action representations were thus robust, specific, and discretized.

First with the 3-T Face/House images, this investigation of the GRL model warranted a wider sample of all learners for the sake of incorporating variability in generalization strategies or lack thereof (Figure 10a, Tables S19, S21, and S22; see S21 and S22 for summary). Crucially, those participants who did learn well were not necessarily taking advantage of the opportunities to generalize. With regard to the primary factor of state generalization, the network implicated in the interaction effect between RPE signals and the strength of generalization was significant at the set level for the same mesostriatal ROIs from the earlier RPE analysis (SVC $p_{FWE} < 0.05$). These state-generalization interactions were aligned with the focal coordinate-based ROI in the SN and also found in the striatum ($p < .005$), including the posterior putamen in the vicinity of the dorsal caudate nucleus (SVC $p_{FWE} < 0.05$) (cf. Doll, Duncan, et al., 2015; Horga et al., 2015; Lee et al., 2014; O'Doherty et al., 2003; Tricomi et al., 2009; Wunderlich et al., 2012)

(Figure 10a). These generalization effects applied to the hippocampus as well ($p < .005$, SVC $p_{FWE} < 0.05$), confirming our hypothesis that this region is involved in relaying generalized learning signals to linked value representations.

For the 7-T Color/Motion version (Figures 10b and S7, Tables S20–S22), effects of state generalization were most robustly identified in the nucleus accumbens ($p < .005$, SVC $p_{FWE} < 0.05$) (Figure 10b). Although the hippocampal result for state generalization did not extend across all learners in this case ($p > .005$), signals in the hippocampus among Good learners did yield generalization effects ($p < .005$, SVC $p_{FWE} < 0.05$). The results for generalization additionally comprised uncorrected effects elsewhere within the anatomical ROIs (SVC $p_{FWE} < 0.10$), and these high-resolution findings are explored thoroughly in consideration of their novelty. Regarding state generalization, the loci of interaction effects included the anterior caudate nucleus (SVC $p_{FWE} < 0.10$) as well as both the SN and the ventral tegmental area (VTA) ($p < .005$) (Figure S7a). Effects of action generalization were found elsewhere in both the SN and the VTA ($p < .005$) (Figure S7b).

A control analysis determined that these findings were specific to generalization as opposed to nonspecific successful learning. Interaction effects between RPE signals and the learning rate (i.e., α/τ) (Tables S23–S25; see S25 for summary) were examined to check for overlap with generalization effects, which would suggest a confound if present. There was in fact no such overlap for either the 3-T Face/House or 7-T Color/Motion results ($p > .005$). Instead, both data sets confirmed the dorsal caudate nucleus as a site where RPE signals are more directly related to learning performance ($p < .005$), replicating previous findings (Colas et al., 2017; Schönberg et al., 2007).

3 | DISCUSSION

Supported by multisite, multifield fMRI in conjunction with computational modeling of both behavioral and neural dynamics, the findings herein have confirmed the merits of the GRL model as representative of a class of RL models obscuring the boundary between model-free caching and model-based inference. This conceptual ambiguity suggests a false dichotomy in the notion of a unidimensional spectrum between these antipodes with ostensible mutual exclusivity; putative roles for dopamine are also complicated by such ambiguity (Botvinick, 2012; Bromberg-Martin et al., 2010; Collins & Cockburn, 2020; da Silva & Hare, 2020; Delgado & Dickerson, 2012; Doll et al., 2012; Eckstein & Collins, 2020; Gardner et al., 2018; Langdon et al., 2018; Nakahara, 2014; Nakahara & Hikosaka, 2012; O'Doherty, 2012; Sadacca et al., 2016; Schultz, 2013). For a structured but challenging learning task that lends itself to implicit generalization with a cognitive map, augmenting the classical RL framework (Sutton & Barto, 1998) with associative and discriminative forms of state and action generalization improved the exposition of human performance at the rigorous individual level—here including idiosyncrasies in generalization. Rather than the unambiguously model-based approaches of the SPE or the HMM that proved less effective here, the intuition of the GRL model parsimoniously remains within the

established bounds of RL and its fundamental RPE signal that is integral to computational analysis of neurophysiology in mesostriatal and corticostriatal circuits. Moreover, this work marks a juxtapositional demonstration of the potential of high-field fMRI for these particular signals and neural systems—especially with respect to the dopaminergic midbrain (cf. Colzoli et al., 2021; de Hollander et al., 2017; Fontanesi, Gluth, Rieskamp, et al., 2019).

Our tripartite neural model—representing interrelated value, decision, and learning signals in parallel—stands among the novel technical and computational contributions made here. Guided by precedents for classic RL (Colas et al., 2017), dynamic RPE signals derived from the GRL model were localized within the dopaminergic midbrain and both ventral and dorsal areas of the striatum. Dissociable value signals from the GRL model could be identified simultaneously in other subregions of the ventral striatum as well as in vmPFC and PCC, amounting to all of the areas hypothesized with meta-analytic priors (Bartra et al., 2013; Clithero & Rangel, 2014). Neural correlates of RT were also controlled for and validated in MFC (Yarkoni et al., 2009) as a proxy for decision-making signals. Furthermore, effects of state and action generalization per se were evident in modulation of RPE signaling in the basal ganglia that could only be accounted for with the GRL model as opposed to basic RL. These interaction effects reflect relaying of RPE signals to representations of other states and actions rather than merely that of the state-action pair experienced at a given moment. The hippocampus was also identified as a hub for mediating this generalization across representations that here would correspond to motor and premotor cortex or visual cortex, including the FFA, the PPA, V4, and MT.

Generalization of knowledge is a ubiquitous cognitive phenomenon that is essential for processing the plethora of different stimuli that organisms encounter (Bush & Mosteller, 1951b; Ghirlanda & Enquist, 2003; Harlow, 1949; Shepard, 1957, 1987; Tenenbaum & Griffiths, 2001; Tversky, 1977), but the broad concept of generalization can manifest itself in myriad different ways depending on the situation. For example, whereas the present paradigm contrasts associative generalization against discriminative generalization among temporally interleaved states and actions that are explored in parallel, alternative paradigms have instead focused on more straightforward associative generalization from familiar or proximal stimuli to novel or distal stimuli of varying apparent similarity (Collins & Frank, 2013; Doll, Duncan, et al., 2015; Doll, Shohamy, & Daw, 2015; Gershman, 2017; Gershman & Niv, 2015; Kahnt et al., 2012; Kahnt & Tobler, 2016; Karagoz et al., 2022; Kool et al., 2016, 2017, 2018; Lesaint et al., 2014; Norbury et al., 2018; Stojić et al., 2020; Tomov et al., 2018; van Dam & Ernst, 2015). The discriminative generalization of GRL is distinguished from such feature-based generalization in the arbitrary mapping of abstract states, thus going beyond simply generalizing across common perceptual features of cues or linked outcomes without state discrimination. Here, the state category is not processed as a unitary representation but rather as a set of representations with a cognitive map (whether implicit or explicit) that discriminates and determines relations within the set as defined by a hierarchical metastate. Another distinction can be drawn between generalized

information and counterfactual information that is made explicit with directly observed feedback rather than inferred from assumptions of interdependence, producing contextual effects such as fictive regret or framing (Camille et al., 2004; Coricelli et al., 2005; D'Ardenne et al., 2013; Li & Daw, 2011; Lohrenz et al., 2007; Montague et al., 2006; Palminteri et al., 2015; Palminteri, Lefebvre, et al., 2017; Pischedda et al., 2020). For this and other reasons (see below), the present label of “generalized RL” is more precise than “counterfactual RL”, for example. The GRL model aims toward broader theoretical advancement for a still-nascent perspective on quasi-model-based extensions of model-free RL, and two dichotomies are formalized in doing so: state versus action generalization and associative versus discriminative generalization, where in this case the latter translates to suboptimal overgeneralization (or conflation) versus optimal inference or pseudoinference.

Not only these dichotomies but also the particular delta-learning rules of the two-dimensional GRL model distinguish it from previous modifications of model-free RL. Often arrived at without the due diligence of model comparison, some modifications have simply yoked value representations—for example, $Q_t(s_t, a_1) \equiv -Q_t(s_t, a_2)$ —or otherwise incorporated only one type of generalization (Aquino et al., 2020; Balcaras & Womelsdorf, 2016; Ballard et al., 2019; Baram et al., 2021; Charpentier et al., 2020; Collette et al., 2017; Daw & Shohamy, 2008; Gläscher et al., 2009; Hampton et al., 2007; Hauser et al., 2014, 2015; Lesage & Verguts, 2021; Liu et al., 2021; Matsumoto et al., 2007; Mattar & Daw, 2018; Reiter et al., 2017; Vinckier et al., 2016; Wimmer et al., 2012; Zaki et al., 2016). Moreover, such models are often formulated without parameterization (e.g., $g_A = -1$) or with a second, counterfactual RPE inverting the only outcome (i.e., $r' = -r$ or $r' = 0$ for $r > 0$) in parallel—and, by extension, multiple RPEs as required—as opposed to the current algorithmic scheme of GRL with weighted duplications of the original RPE signal to be relayed to parallel representations of estimated values. The assumption of an inverted counterfactual outcome is not universally applicable and can also create scaling problems for value signals, including disproportionate RPEs as well as even illogical probability estimates ($P < 0$ or $P > 1$) or negative value estimates despite strictly positive outcomes (or vice versa). This issue is especially problematic for modeling that is less abstract in its application to an actual neural system. Another issue is counterfactual learning via separate hypothetical RPE signals each subtracting their respective reward predictions, which is less tractable for modeling than the present implementation based on a parameterized heuristic with relayed duplication of a single RPE signal: Relaying—or perhaps even multiplexing (cf. Nakahara, 2014; Nakahara & Hikosaka, 2012)—is less computationally demanding and more parsimonious. (A multiplexed signal in this context would additionally specify how the RPE is to be generalized.) It is less plausible that any number of hypothetical RPE signals could be distinguished in the brain in parallel with assumptions of unique RPE signals for each of the states or actions updated per single action performed—and particularly implausible for an action space that is more continuous rather than discrete. The GRL model therefore eschews a true “counterfactual RPE” in favor of a “generalized RPE” (but still can be regarded as a version of counterfactual

learning). This formulation is readily scaled up for environments with arbitrary complexity in the numbers of state-action pairs or category-state-action triplets.

This efficient approach is analogous to the “TD(λ)” eligibility trace (Dayan, 1992; Dayan & Sejnowski, 1994; Klopf, 1972; Sutton, 1988; Sutton & Barto, 1981, 1998) that forgoes separate RPEs for antecedent events in favor of more conservatively duplicating, reweighting, and relaying the current RPE back in time with decay along the memory trace (Figure 2). This perspective of TD(λ) as temporal generalization could consider it as a particular form of associative generalization across linear time, which evokes the temporal spread of the law of effect (Thorndike, 1911, 1933). In addition to the dimension of time, the RPE in this “GRL(λ)” model is generalized across dimensions in the abstract space of state and action representations with nonlinear temporal mapping (cf. Jocham et al., 2016). This topological space could be encoded in nonspatial cognitive maps analogous to the location-based spatial maps (Moser et al., 2008; O'Keefe & Nadel, 1978) represented in the medial temporal lobe, including the hippocampus and entorhinal cortex (Ballard et al., 2019; Baram et al., 2021; Behrens et al., 2018; Bernardi et al., 2020; Cazé et al., 2018; Daw & Shohamy, 2008; Gerraty et al., 2014; Liu et al., 2019, 2021; Mattar & Daw, 2018; Momennejad et al., 2018; Park et al., 2020; Wimmer et al., 2012; Wimmer & Shohamy, 2012). Such abstract cognitive mapping (Tolman, 1948) follows from the intrinsic topology of mental representations as postulated in field theory (Lewin, 1935, 1936). Simulated replay of experienced (or even hypothetical) events by the hippocampus is a potential mechanism for recapitulating task-relevant information (Cazé et al., 2018; Eldar et al., 2020; Gershman et al., 2014; Kurth-Nelson et al., 2016; Liu et al., 2019, 2021; Mattar & Daw, 2018; Momennejad et al., 2018; Schuck & Niv, 2019; Wimmer et al., 2020), which for the present purposes could contribute to the associations underpinning generalization. This hippocampal replay is reminiscent of the model-free but quasi-model-based “Dyna” architecture from machine learning that approximates model-based dynamic programming as indirect RL with quasi-inferential iterations of simulated experiences (Sutton, 1990, 1991).

The algorithm described here follows a rising trend toward model-based alternatives to the model-free RL framework, including hybrid models that integrate multiple learning systems (Daw et al., 2005; Doll et al., 2012; O'Doherty et al., 2017, 2021). With the simplest case of two systems receiving the most examination, the theoretical dichotomy of model-free and model-based processes is analogous to an extent with that between habitual (Pavlov, 1927; Thorndike, 1898, 1911) and goal-directed (Tolman, 1948) learning. The most commonly studied domain of model-based inference has typically been characterized with a modular system engaged in explicit forward planning of future behaviors in parallel with the caching of model-free associations (Charpentier et al., 2020; Daw et al., 2005, 2011; Gläscher et al., 2010; Lee et al., 2014). For example, dynamic programming can achieve optimal goal-directed behavior in a multi-step Markov decision process (MDP) with the learning of transition functions for states and state-action pairs (Bellman, 1957; Sutton & Barto, 1998), which can be arrived at with computation of another

type of SPE analogous to the RPE (Gläscher et al., 2010; Lee et al., 2014). However, the broader model-based umbrella can also encompass cognitive maps and certain mechanisms for generalization in learning, including the HMM (Ghahramani, 2001; Hampton et al., 2006; Prévost et al., 2013) and other Bayesian processes (Tenenbaum & Griffiths, 2001) as well as the novel formulation of the SPE developed here (with an “MPE” for metastates). Such generalization represents a potential domain of overlap between model-based and model-free processes (Bromberg-Martin et al., 2010; Doll et al., 2012; Doll, Duncan, et al., 2015; Doll, Shohamy, & Daw, 2015; Hampton et al., 2006, 2007; Karagoz et al., 2022; Kool et al., 2016, 2017, 2018; Liu et al., 2021; Mattar & Daw, 2018; O’Doherty, 2012; Sadacca et al., 2016; Wimmer et al., 2012; Wunderlich et al., 2011).

As GRL features an implicit model for generalization while prioritizing parsimony and computational efficiency, this scheme—by way of analogy to Dyna—is not neatly encapsulated by either extreme of the model-free/model-based dichotomy. GRL thus also joins the ranks of the successor-representation algorithms that operate with analogous ambiguity in shortcut solutions based on a compressed transition function, which would be more applicable to a learning task with multiple steps per episode (or trial) (Akam et al., 2015; Dayan, 1993; Momennejad et al., 2017; Russek et al., 2017, 2021). Likewise, whereas GRL frugally accounts for generalization across states and actions within a task, by extension, heuristic algorithms based on reward-predictive state abstractions have been proposed for generalization across tasks, which can be represented by unique transition functions (Franklin & Frank, 2018; Lehnert et al., 2020; Li et al., 2006).

As the GRL model forgoes supplanting model-free RL altogether, so too does it forgo complementing a model-free system with a model-based system operating in parallel (cf. Doll, Duncan, et al., 2015; Doll, Shohamy, & Daw, 2015; Karagoz et al., 2022; Kool et al., 2016, 2017, 2018)—instead opting for quasi-model-based augmentation of model-free RL that is still effectively characterized by a single system. From the perspective of control theory, an HMM (Ghahramani, 2001) can provide an optimal Bayesian solution to this generalization problem with a fully model-based approach to structural inference (Hampton et al., 2006; Prévost et al., 2013), but this avenue entails assumptions of more complex computations as well as ambiguity concerning the physical implementation of the implied neural mechanisms (cf. Gläscher et al., 2009; Hampton et al., 2007); the latter can be an obstacle to achieving comprehensive triangulation across levels of analysis (Marr, 1982). Although engagement of multiple systems for a model-free and model-based hybrid remains within the realm of possibility, going down this route of additional moving parts with two systems—let alone in excess of two—is even more problematic in this regard (cf. Daw et al., 2011). Presently, GRL outperformed dual-systems alternatives despite whatever viability they could have. In addition to the virtue of Occam’s razor on the theoretical side of parsimonious modeling (Myung, 2000), there are practical advantages for fitting and interpretability with the simpler RL-based approach in settings where fully model-based learning is less essential or even counterproductive for a dynamic environment. That is, a primarily model-free strategy is often sufficient for at least near-optimal performance, and humans (like other animals) often fail to achieve optimal performance anyway, as was evident here and in another study (Aquino

et al., 2020). These benefits extend to modeling of not only behavior but also neurophysiology, where a parsimonious model grounded in well-defined concepts can provide a stable foundation with utility such as for interpreting performance of varied tasks or for identifying nodes in relevant networks (Bassett et al., 2018; Gerraty et al., 2018; Mattar et al., 2018). In this case, subjective value and the RPE naturally fill roles as part of a trichotomy of value, decision, and learning signals in the brain that collectively function as the interface between sensory input and motor output.

This modeling sets the stage for further inquiry concerning how arbitration among strategies for generalization might be implemented; here lies an analogy with reliability-based arbitration among modular model-free and model-based systems to integrate information across a “mixture of experts” (Charpentier et al., 2020; Daw et al., 2005; Lee et al., 2014; O’Doherty et al., 2017; O’Doherty et al., 2021) as in machine learning (Hamrick et al., 2017; Jacobs et al., 1991; Masoudnia & Ebrahimpour, 2014; Yuksel et al., 2012). Instead of tracking absolute prediction errors or entropy in updates of cached value functions or modeled transition functions, the “experts” for generalization would be concerned with tracking regularities or irregularities across inputs for the structural models embedded in such functions (as well as tracking task demands warranting effort) (cf. Hampton et al., 2006; Karagoz et al., 2022; Kool et al., 2017, 2018; Lehnert et al., 2020; Liu et al., 2021; Mattar & Daw, 2018; Prévost et al., 2013; Schulz et al., 2018, 2020; Wu et al., 2019; Wu, Schulz, Garvert, et al., 2018; Wu, Schulz, Speekenbrink, et al., 2018; Wunderlich et al., 2011). However, modeling such arbitrated metalearning for the present experiment is precluded by certain practical limitations—in particular, the issue of both forms of discriminative generalization always being optimal strategies and hence not being modulated with sufficient variability. The GRL model in its current form does not distinguish between prior assumptions about generalizable structure and learned information about structure acquired through serial observations. Feasibly translating the static generalization effects of the current GRL model to dynamical generalization processes will require an experimental paradigm with dynamic structure more directly catered to manipulating cognitive models for generalization—for example, including alternation across correlation, anticorrelation, and independence. Nevertheless, the presently static generalization parameters suffice as an initial proof of concept for such extensions of RL and in particular both associative and discriminative generalization across both states and actions.

This endeavor has justified the GRL model as a viable and practical tool in a growing model space that need not be limited to purely model-free learning and purely model-based learning. These results localize a modular network of brain regions that orchestrate evaluation and value-based learning and decision making in a setting characterized by generalizable patterns across both states and actions. By identifying the nodes of a network mediating reinforcement learning and concomitant generalization to link representations of stimuli or motor responses within a cognitive map, this (computational) model-based mapping lays the groundwork for further investigation of network dynamics (Bassett et al., 2018; Gerraty et al., 2018; Mattar et al., 2018) with the potential to yield yet more comprehensive understanding of the causal chain of information flow between sensation and action in a reward-based environment that is noisy and dynamic but also predictably structured.

4 | METHODS

4.1 | 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 and 7-T versions of the study, respectively. This collaborative multisite study was conducted at six separate facilities for magnetic-resonance imaging (MRI), such that 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. Participants provided informed written consent according to protocols approved by the respective Institutional Review Boards of each scanning site—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. Upon completing the study, participants were paid \$10 for minimizing head movement plus the amount of money earned within the task.

4.2 | Experimental procedures: 3-T Face/House version

Shown in Figure 1 is a schematic of the hierarchical reversal-learning task that includes outcome probabilities for every combination of state and action within one of 12 blocks defined by said probabilities. (A complete session of 12 blocks is detailed in Figure 7.) 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 an image of a face or a house appearing against a white background subtending $8.1^\circ \times 8.1^\circ$ of visual angle at the center of the display—first flanked by two white arrows to the left and right each subtending $1.0^\circ \times 8.1^\circ$ and centered at an eccentricity of 4.9° . The participant was allotted 2 s to respond to this two-armed bandit by pressing one of two buttons with the corresponding index finger of either the left or right hand. To confirm the response while minimizing eye movement, the arrow corresponding to the nonchosen action was removed from the display between the time of response and stimulus offset. A fixed interstimulus interval (ISI) of 3 s separated the cue and the outcome. In consideration of the sensitivity of a TD learning algorithm to the timing of outcomes (McClure et al., 2003; O'Doherty et al., 2003; O'Doherty et al., 2004; Sutton, 1988; Sutton & Barto, 1998), jitter—otherwise typical of rapid event-related designs in functional MRI (fMRI)—was forgone with the ISI in favor of a design that induced stable prediction-error signals.

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 a black dollar sign against a white background again subtending $8.1^\circ \times 8.1^\circ$ for 1 s, whereas a scrambled dollar sign signified an absence of monetary reward for that trial. This scrambled image was generated by randomly rearranging segments of the dollar sign as a regular 8×15 grid. Only a white fixation cross subtending $0.7^\circ \times 0.7^\circ$ of visual angle was presented at the center of a black background throughout the ISI and the intertrial interval (ITI). This fixation cross also remained in the foreground of the display with a black outline during stimulus presentation. The duration of the jittered 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.

Representing each active state, four new cues were assigned randomly every run with two pairs of images each respectively drawn from two state categories. In the 3-T version of the experiment, these categories were faces and houses, which share common low-level visual features as a control. Face stimuli were extracted from the Chicago Face Database (Ma et al., 2015), which also includes subjective ratings of the stimuli along various dimensions. A set of eight face images were selected for depicting an adult male who was consistently classified in the “White” ethnic group ($M = 97.5\%$, $SD = 2.0\%$) and rated as neither especially attractive nor especially unattractive (Likert scale [1, 7]: $M = 3.55$, $SD = 0.23$). All portraits were intended to display a neutral facial expression. These selection criteria minimized the potential for hedonic evaluation of the arbitrary stimuli themselves to interfere with experimental manipulations of value-based associations based on rewards in the task (Chien et al., 2016). In keeping with these controls, all images were converted to grayscale. House stimuli were extracted from the DalHouses database (Filliter et al., 2016), which included subjective ratings of facial pareidolia and other attributes. A set of eight house images were selected for being rated as minimally facelike (Likert scale [1, 7]: $M = 2.22$, $SD = 0.10$) and being distinctive relative to the rest of the set. As the human brain is endowed with innate expertise for recognizing faces but not houses (Kanwisher, 2000), the face stimuli were selected to be homogenous while the house stimuli were instead selected to maximize the heterogeneity of the set.

Rather than sheer randomness, which especially limits 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 made for 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 $3 \times 2 \times 4$ design for the 12 blocks, the 3×2 and 3×4 crosses were fully counterbalanced while the 2×4 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 actions within a state and between states within a category. The categories were independent of each other without any such structured pattern between them.

The first condition (“3” in the $3 \times 2 \times 4$ design), having three possibilities also counterbalanced within a run, 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 (“2”), having two possibilities partially counterbalanced with a 2:1 ratio within a run, 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 (“4”), having four possibilities, concerned the mapping of a category’s reward probabilities to actions, such that the two states (“A” and “B”) within a category always symmetrically provided rewards for opposite actions. This complementarity between states within a category was designed to reveal state generalization. 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.

Between blocks, the design was constrained for a single remapping—that is, reversals of rewarded actions within only one category—to mark the onset of a new block within a run. The two categories were remapped in turn in a random order counterbalanced across runs, such that each category had 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.

Stimuli were projected onto a screen that was viewed with an angled mirror in the MRI scanner. The viewing distance was 100 cm in the case of the Caltech sample, which served as the basis for the approximate stimulus sizes reported here, but there was slight variability in stimulus sizes across laboratories. The display was presented with a resolution of 1024×768 pixels and a refresh rate of 60 Hz. The primary stimuli had a resolution of 375×375 pixels. The interface was programmed with MATLAB (MathWorks) and the Psychophysics Toolbox (Brainard, 1997).

4.3 | Experimental procedures: 7-T Color/Motion version

Conducted in parallel, the second version of the experiment was mostly matched to the first but was not entirely identical. Only differences between versions are emphasized in this section.

This 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 were not replaced every run as with the 3-T version’s faces and houses. Although the two pairs of stimuli comprising the two categories remained constant across the entire session, the factorial design of the 3-T version was preserved such that the reward probabilities for these constant states still rotated as before. The 7-T version was fully counterbalanced as before, and the constant cues allowed for even further counterbalancing such that each cue preceded exactly one quarter of the available rewards in a session.

The color stimuli were flickering dot arrays that alternated between two colors for each state. One stimulus alternated six times between red dot arrays and green arrays at a rate of 6 Hz, and the other similarly alternated between blue and yellow dots. These color stimuli were essentially arranged as static frames of the motion category’s random-dot kinetograms (Newsome & Paré, 1988). Apparent motion was generated by displacement of the dots in a consistent direction every frame with 100% motion coherence. The two states in this category were represented with upward or downward motion, respectively. The speed of displacement was 2.1° per second.

For both color and motion stimuli, unique dot arrays were randomly generated with every trial. These arrays contained over 100 square dots randomly positioned against a black background. The array was framed by a gray square subtending $4.1^\circ \times 4.1^\circ$ at the center of the display—first flanked by two gray arrows each subtending $0.5^\circ \times 4.1^\circ$ and centered at an eccentricity of 2.5° . If the participant failed to respond in time, the nonrewarded outcome appeared immediately as the white fixation cross subtending $0.4^\circ \times 0.4^\circ$ turned gray.

4.4 | Data acquisition: 3-T Face/House version

For the first version of the experiment, MRI data were collected with a common set of protocols across five sites housing 3-tesla Magnetom Prisma scanners (Siemens Medical Solutions, Malvern, PA)

equipped with 2-channel body-transmit and 32-channel head-receive coils. The first structural volume covered the whole brain and was acquired to guide subsequent functional imaging with a single-inversion T1-weighted (T1w) 3-dimensional (3D) magnetization-prepared rapid gradient-echo (MPRAGE) sequence that had the following parameters: repetition time (TR): 2400 ms, echo time (TE): 2.32 ms, inversion time (TI): 800 ms, RAGE flip angle (FA): 10°, in-plane GRAPPA acceleration factor (R): 2, voxel: 0.9 mm isotropic, field of view (FOV): 187 × 230 × 230 mm. A second structural volume was acquired after the experiment with a T2-weighted (T2w) 3D SPACE (“sampling perfection with application-optimized contrasts using different flip-angle evolutions”) sequence that had the following parameters: TR: 3200 ms, TE: 564 ms, FA: variable, R: 2, voxel: 0.9 mm isotropic, FOV: 187 × 230 × 230 mm.

During the experiment, functional images were acquired from the whole brain using a blood-oxygen-level-dependent (BOLD) contrast with a T2*-weighted gradient-echo echo-planar imaging (EPI) sequence (Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota) featuring both in-plane GRAPPA (“generalized autocalibrating partially parallel acquisitions”) (Griswold et al., 2002) and multiband slice excitation (Feinberg & Setsompop, 2013; Moeller et al., 2010) and having the following parameters: TR: 1120 ms, TE: 30 ms, FA: 54°, multiband acceleration factor (M): 4, R: 2, voxel: 2.0 mm isotropic, FOV: 144 × 192 × 192 mm. Off-resonance distortion correction was based on phase-encoding polarity-reversed spin-echo EPI image pairs with geometry, acceleration, and EPI echo spacing all matched to the BOLD fMRI series (TR: 5130 ms, TE: 41.4 ms, FA: 90°, voxel: 2.0 mm isotropic, FOV: 144 × 192 × 192 mm). The session consisted of four functional runs each having a duration of 17.7 min and each preceded by field maps.

Peripheral cardiac and respiratory signals were recorded during scanning by way of scanner-integrated wireless sensors. The pulse sensor was attached to the ring finger of either the left hand or the right hand, such that this factor was counterbalanced across subjects. The pneumatic sensor was secured under a strap to measure external displacement of the lungs.

4.5 | Data acquisition: 7-T Color/Motion version

For the second version of the experiment, MRI data were collected at a single site using a 7-tesla Siemens Magnetom Terra scanner equipped with a single-channel head-transmit volume coil and a 32-channel head-receive coil. In light of the tradeoff between the signal-to-noise ratio (SNR) or the contrast-to-noise ratio (CNR) and either spatial or temporal resolution, high-field neuroimaging allows for a superior SNR and CNR (De Martino et al., 2018; Dumoulin et al., 2018; Torrisi et al., 2018; Uğurbil, 2018) that could be relied upon here to achieve higher spatial resolution. In the interest of maximizing spatial resolution at 7 T, temporal resolution for the EPI sequence was also compromised somewhat relative to the 3-T protocol, but simultaneous multislice acquisition still enabled viable temporal resolution. Enhancing the volumetric resolution by a factor of 4.6, this approach boasted more precise discernment of

mesencephalic nuclei (Eapen et al., 2011) in particular. Otherwise, the 7-T protocols were matched to the 3-T protocols as closely as possible to allow for direct comparison.

The session again began with a whole-brain structural volume acquired using a dual-inversion T1w 3D “MP2RAGE” sequence (Choi et al., 2019) (TR: 4010 ms, TE: 2.86 ms, T₁: 1050 ms, T₂: 3200 ms, FA₁: 6°, FA₂: 4°, R: 2, voxel: 0.8 mm isotropic, FOV: 179 × 220 × 220 mm). The complementary structural volume was acquired after the experiment with a T2w 3D SPACE sequence (TR: 4270 ms, TE: 315 ms, FA: variable, R: 3, voxel: 0.7 mm isotropic, FOV: 168 × 224 × 224 mm).

During the experiment, functional images were acquired from the whole brain using a BOLD contrast with a T2*-weighted gradient-echo EPI sequence (Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota) (TR: 1960 ms, TE: 22 ms, FA: 45°, M: 4, R: 2, voxel: 1.2 mm isotropic, FOV: 125 × 192 × 192 mm). Off-resonance distortion correction was based on phase-encoding polarity-reversed spin-echo EPI image pairs with geometry, acceleration, and EPI echo spacing all matched to the BOLD fMRI series (TR: 7680 ms, TE: 30.6 ms, FA: 90°, voxel: 1.2 mm isotropic, FOV: 125 × 192 × 192 mm). Cardiac and respiratory signals were again recorded via wireless sensors during scanning.

4.6 | Data preprocessing

Real-valued, signed T1w images generated by the MP2RAGE sequence at 7 T were first masked with Otsu thresholding (Otsu, 1979) of auxiliary magnitude data generated by the same sequence, thereby eliminating background noise in surrounding air. This initial step ensured compatibility with subsequent structural preprocessing as part of a common pipeline applied to both 3-T and 7-T images.

Neuroimaging data were primarily preprocessed using fMRIPrep version 1.2.5 (Esteban et al., 2019). This software package includes elements from the FMRIB Software Library (FSL) v5.0.9 (Centre for fMRI of the Brain, University of Oxford) (Smith et al., 2004), Advanced Normalization Tools (ANTs) v2.1.0 (Avants et al., 2010), FreeSurfer v6.0.1 (Laboratory for Computational Neuroimaging, Athinoula A. Martinos Center for Biomedical Imaging) (Fischl, 2012), and Analysis of Functional NeuroImages (AFNI) v16.2.07 (Scientific and Statistical Computing Core, National Institute of Mental Health) (Cox, 1996)—all compiled with the Nipype interface (Gorgolewski et al., 2011) and often facilitated by the Nilearn toolbox (Abraham et al., 2014).

Rather than utilizing the scanning system's internal bias-field correction, each T1w volume was first corrected for intensity nonuniformity using “N4BiasFieldCorrection” (ANTs) (Tustison et al., 2010). Skull stripping was then performed with “antsBrainExtraction” (ANTs) using the Open Access Series of Imaging Studies (OASIS) template (Marcus et al., 2007). Incorporating information from both T1w and T2w volumes, brain surfaces were reconstructed using “recon-all” (FreeSurfer) (Dale et al., 1999). The previously estimated brain mask was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of cortical gray matter

from Mindboggle (Klein et al., 2017). All images were converted to the common Montreal Neurological Institute (MNI) space (Collins et al., 1994). Spatial normalization to the MNI152-based ICBM 2009c Nonlinear Asymmetric template (Fonov et al., 2009) was performed through nonlinear registration with “antsRegistration” (ANTs) (Avants et al., 2008), combining brain-extracted versions of both the T1w volume and the template. Segmentation of brain tissue into cerebrospinal fluid (CSF), white matter, and gray matter was performed on the brain-extracted T1w volume using FMRIB's Automated Segmentation Tool (FAST) (FSL) (Zhang et al., 2001).

For functional BOLD images, slice-time correction was applied with “3dTshift” (AFNI). Motion correction was applied using Motion Correction with FMRIB's Linear Image Registration Tool (MCFLIRT) (FSL) (Jenkinson et al., 2002). Utilizing field maps, distortion correction and unwarping was performed with an implementation of the TOPUP technique (Andersson et al., 2003) using “3dQwarp” (AFNI). Functional volumes were coregistered to the corresponding structural T1w volume via boundary-based registration (Greve & Fischl, 2009) with 9 degrees of freedom, as implemented by “bbregister” (FreeSurfer). All transformations for motion correction, distortion correction, BOLD-to-T1 coregistration, and T1-to-template coregistration were concatenated and applied in a single step using “antsApplyTransforms” (ANTs) with Lanczos interpolation.

First using only imaging data itself for denoising, dynamics of sources of noise such as head motion, physiological events, and measurement (i.e., scanner) artifacts were estimated by different approaches to generate confound regressors of no interest for the general linear model (GLM). Notably, BOLD signals throughout the brainstem have an intrinsically low SNR and a low CNR and are especially susceptible to physiological artifacts (Barry et al., 2013; Dagi et al., 1999; de Hollander et al., 2015, 2017; Düzel et al., 2009, 2015; Enzmann & Pelc, 1992; Soellinger et al., 2007). The proximity of the pulsatile interpeduncular cistern to the tegmentum further compromises signals of purely neural origin in the key region of the dopaminergic midbrain. To address these issues as well as possible differences in output between the different scanners for multisite fMRI, further extensive efforts were dedicated to eliminating contaminant noise as follows.

Six rigid-body motion parameters—corresponding to three axes for translation and three for rotation—were estimated relative to a reference image and subsequently added to the design matrix. Time series of signals averaged within the CSF mask, within the white-matter mask, or globally across the entire brain mask were included next. Framewise displacement quantified bulk head motion within each functional run (Power et al., 2012, 2014). An index for the rate of signal change across the entire brain was provided with the standardized temporal derivative of root-mean-squared variance over voxels (DVARs) (Power et al., 2012, 2014; Smyser et al., 2010). Initial time points identified as nonsteady states according to global signals were marked with unique indicator variables for each outlier volume. Furthermore, these outliers were omitted from the following denoising procedures.

For temporal high-pass filtering, a discrete cosine transform (DCT) (Ahmed et al., 1974) was employed to detect low-frequency signal drift. Fifteen DCT basis functions were generated as regressors after omitting the aforementioned outliers. The CompCor method for denoising relied on principal-component analysis, and principal components were generated with the two variants of the algorithm—namely, “temporal” (tCompCor) and “anatomical” (aCompCor) (Behzadi et al., 2007). A mask to exclude signals with cortical origins was first obtained by eroding the brain mask so as to ensure it only contained subcortical structures. Six tCompCor components were then estimated with voxels above the 95th percentile for signal variability within the eroded subcortical mask. Another six aCompCor components were estimated within the intersection of the subcortical mask and the union of CSF and white-matter masks in T1w space after projection to the native space of each functional run. Employing probabilistic spatial independent-component analysis (ICA) as implemented by Multivariate Exploratory Linear Decomposition into Independent Components (MELODIC) (FSL) (Beckmann & Smith, 2004), the “aggressive” ICA-based strategy for Automatic Removal of Motion Artifacts (ICA-AROMA) (Pruim et al., 2015) was utilized to distinguish signal and noise components with dimensionality constrained to a maximum of 200 components.

Additional preprocessing was performed outside of fMRIPrep using Statistical Parametric Mapping (SPM) v12.7219 (Wellcome Centre for Human Neuroimaging, University College London) (Friston et al., 1995). Spatial smoothing was a final step, convolving functional images with an isotropic Gaussian kernel having a full width at half maximum (FWHM) of 6 mm for the 3-T data set. As the aim of the 7-T protocol was to maximize spatial resolution, the FWHM parameter was reduced to 2 mm for 7-T data and thus preserved the fine granularity critical for detecting mesencephalic signals (Chase et al., 2015; de Hollander et al., 2015).

Moreover, the PhysIO toolbox (Kasper et al., 2017) was used to produce confound regressors derived not with imaging data but rather with peripheral cardiac and respiratory recordings. The retrospective image correction (RETROICOR) method (Glover et al., 2000) generated third-order Fourier expansion of the cardiac phase (i.e., 6 terms for sine and cosine functions), fourth-order expansion of the respiratory phase (8 terms), and first-order expansion of cardiorespiratory interactions (4 terms) (as parameterized optimally in Harvey et al., 2008).

4.7 | Computational modeling: Generalized reinforcement learning

As a quasi-model-based extension of model-free “reinforcement learning” (RL) (Bush & Mosteller, 1951a; Rescorla & Wagner, 1972; Sutton & Barto, 1998) with the temporal-difference (TD) prediction method (Dayan, 1992; Dayan & Sejnowski, 1994; Sutton, 1988), this “generalized reinforcement learning” (GRL) model introduced the dichotomies of associative versus discriminative generalization and state versus action generalization within the “critic/Q-learner”

(CQ) model (Colas et al., 2017) (Figure 2). The CQ model integrates the “critic” component of the “actor/critic” model (i.e., state-value learning) (Barto et al., 1983, 2021; Sutton, 1984; Witten, 1977) with the Q-learning model (i.e., action-value learning) (Watkins, 1989; Watkins & Dayan, 1992) for passive and active states, respectively. If it were instead a question of one model or the other for a paradigm such as this having few discrete and constant actions, the Q-learning model typically provides more accurate fits to behavior (Colas et al., 2017; Hampton et al., 2006; O’Doherty et al., 2004); the actor/critic model is instead ideal for a broad, continuous, or dynamic action space. Although further hybridization of the two algorithms has been demonstrated (Colas et al., 2017), the “actor” module of the proper “actor/critic/Q-learner” (ACQ) model was omitted here because adding a costly free parameter for this module is less essential for a task with only one cue and two possible actions per trial; in any case, this additional complexity was beyond the scope of the present study despite the otherwise relevant handling of passive and active states. Here, this richer account of internal decision variables—one that goes beyond what is immediately evident in behavior—facilitated not only theory but also the interpretability of the neuroimaging analysis for triple dissociation of value signals, RPE signals, and decision signals modulated by value.

By design, the model is scalable for arbitrary numbers of hierarchically organized actions, states, and state categories. Yet, for clarity, the equations herein are not written in their general form (see Colas et al., 2017, for CQ(λ) sans generalization) but rather are tailored to only what is applicable for the present paradigm. The CQ model employs two variants of the reward-prediction error (RPE) to learn value-based associations—namely, the state-value-prediction error (SVPE) and the action-value-prediction error (AVPE). A more precise label for the GRL model postulated here could be the “generalized critic/Q-learner” (GCQ) model, but generalized Q learning was the primary mechanism under scrutiny. Whereas such a critic module would feature only state generalization, the “generalized Q-learner” (GQL) module features both state and action generalization.

To begin with, only the preparatory state of the ITI was represented by the CQ model’s critic module as a passive state s_0 . As representing priors in the absence of previous associations would entail some kind of internal model, a naïve model-free agent initializes the value of this novel state $V_t(s_0)$ to zero (Li et al., 2011):

$$V_0(s_0) = 0$$

The Q-learner module is instead concerned with the active states. The beginning of a run marks initialization of action values $Q_t(s, a)$ for all novel state-action pairs—again at zero:

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

Upon transitioning from the preparatory state to an active state, an SVPE δ_t^V is computed as the difference between cached values per a TD algorithm. Despite not necessarily being relevant for behavior at the moment of exposure, passive states are tracked automatically because behavioral relevance can be unpredictable in the real world

(Colas et al., 2017). For the sake of parsimony, the relevant input here is the prespecified action value rather than an additionally posited state value that could be represented in parallel by the critic module (cf. Colas et al., 2017). There was only one opportunity for action per episode in the present paradigm, so as far as fitting behavior is concerned, the “off-policy” Q-learning method could not be distinguished from an “on-policy” alternative such as the state-action-reward-state-action (SARSA) method (Rummery & Niranjan, 1994). The former computes an RPE using the maximal value across subsequently available actions, whereas the latter computes an RPE using the value of the action actually chosen according to the current policy. Distinguishing these particular algorithms would require at least one additional step with an active state per episode. Yet, as the original standard for an action-value-learning algorithm, Q learning was assumed for neural modeling without further consideration of the SARSA model. Additionally, the standard discount factor γ was omitted here (i.e., $\gamma = 1$) inasmuch as only one reward could be delivered after a constant delay within episodic trials, leaving this reduced delta-learning rule:

$$\delta_t^V = \max_a Q_t(s_{t+1}, a) - V_t(s_0)$$

The value of the preparatory state is updated in turn with a fitted learning rate α (for $0 \leq \alpha \leq 1$) as follows:

$$V_{t+1}(s_0) = V_t(s_0) + \alpha \delta_t^V$$

Upon transitioning from an active state to an outcome state, an AVPE δ_t^Q is determined by the discrepancy between the current action-value estimate $Q_t(s_t, a_t)$ and the reward (or lack thereof) r_{t+1} presented in the binary outcome state:

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

As with any standard RL model, the value of the chosen state-action pair is updated accordingly once the outcome has been processed. The learned information is assumed to be integrated immediately, which would be an optimal use of the time preceding the next trial:

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

Considering that the reward magnitude is fixed for this paradigm, state values and action values effectively correspond to the probability of reward. To prevent duplicated and relayed prediction errors from producing an illogical expected value for probabilistic outcomes (i.e., $0 \leq P \leq 1$), the function $f(x)$ constrains state and action values between zero and unity as an ad-hoc solution for this case where probability is equivalent to value. Inasmuch as a guaranteed improvement in fit in the absence of this constraint would be uninterpretable here, it is not a possibility that is considered for now: Probability estimates above unity or below zero would be meaningless as probabilities per se, and the latter would also correspond to negative valence despite an absence of punishment. Although reference dependence and normalization are mechanisms of relevance to value-based

learning (Carandini & Heeger, 2012; Kahneman & Tversky, 1979; Palminteri & Lebreton, 2021; Rangel & Clithero, 2012), the present paradigm is not suitably amenable to these complexities. When applied to the (computational) model-based neuroimaging analysis, these simulated signals have substantial implications for the interpretation of value signals in the brain, which should be maximized with certain reward and range from neutral to appetitive rather than including anything in the aversive range. The x here refers to a transformation for an updated value estimate:

$$f(x) = \max\{0, \min\{1, x\}\}$$

With the addition of the “TD(λ)” eligibility trace (Dayan, 1992; Dayan & Sejnowski, 1994; Klopf, 1972; Sutton, 1988; Sutton & Barto, 1981, 1998), this “CQ(λ)” model learns more rapidly with credit assignment across serial events. The eligibility trace of the TD(λ) prediction-error signal weights updates prior to the most immediate one according to the eligibility parameter λ (for $0 \leq \lambda \leq 1$) as the base (i.e., inverse decay rate) of an exponential function modulating the learning rate α . With discretely episodic paradigms such as the present one, the eligibility trace only propagates back to the onset of the trial. Owing to this temporal generalization, the preparatory state is updated by the AVPE as well:

$$V_{t+1}(s_0) = f(V_t(s_0) + \lambda \alpha \delta_t^Q)$$

Thus far, the CQ model has been described in its original form. 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 (Barraclough et al., 2004; Ito & Doya, 2009; Kato & Morita, 2016; Morita & Kato, 2014; Toyama et al., 2017, 2019). (There are intriguing parallels in the mathematics of value decay and counterfactual updating for nonencountered representations that remain to be investigated elsewhere.) In contrast to previous RL models, the GRL (or GQL) model introduced here additionally applies a common AVPE signal to learning of other state-action pairs belonging to the same category as the current state. 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 action and the right action, respectively:

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

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 choice 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 resolves the potential nonidentifiability issue of multiplied free parameters for generalized delta learning. More importantly, this constraint reflects the assumption—one shared with TD(λ)—that here generalized RPE signals would not be relayed with greater gain than the original RPE signal but rather lesser or equal gain. (In a different setting, this assumption might be relaxed under the appropriate circumstances.) As with state generalization, this equation is analogous to the previous one for the temporal generalization of the TD(λ) eligibility trace:

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

Likewise, with only two states per category, state generalization entails an analogous formula where—in addition to the encountered state s_t —the other, complementary state within the category s'_t receives a 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 here, overgeneralization across states within a category can be detected. That is, the agent could incorrectly operate as if the category itself is a unitary state or at least partially conflate exemplars within a category. As the present paradigm is characterized by anticorrelational linkage between states within a category, a negative sign for g_S produces correct discriminative generalization, while a positive sign for g_S produces incorrect associative overgeneralization:

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

As an intuitive constraint for the 7-parameter model, the two factors of action generalization and state generalization interact multiplicatively to also update the complementary action for the complementary state. (This assumption in lieu of a third generalization parameter is also pragmatic in the interest of avoiding overfitting here, but it does not necessarily apply universally.) In the ideal 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:

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

Although the preceding constraint was hypothesized to be an appropriate one here, the possibility of an unconstrained interaction term as part of an 8-parameter model was also considered and tested. Yet, in keeping with the initial constraint of only negative action

generalization (i.e., $-1 \leq g_A \leq 0$), the partial constraint that the two updates for the complementary state could not share a common (non-zero) sign remained such that the interaction term was not determined by a wholly independent free parameter. Rather, a more general version of the preceding equation (from the nested case of $g_{SA} = g_A$) includes the third factor of interactive state-and-action generalization g_{SA} (for $-1 \leq g_{SA} \leq 0$) as follows:

$$Q_{t+1}(s'_t, a'_t) = f\left(Q_t(s'_t, a'_t) + g_S g_{SA} \alpha \delta_t^Q\right)$$

These learned action values serve as inputs to a probabilistic action-selection policy $\pi_t(s, a)$ characterized by the Boltzmann-Gibbs softmax model as a discriminative (rather than generative) model of decision making (Luce, 1959; Shepard, 1957; Sutton & Barto, 1998). The approximation of a softmax (with perfect subtraction between two alternatives) does have limitations in accounting for decision-making processes in an actual brain (Colas, 2017), but this component can suffice for the present purposes as a standard assumption for learning models. The choice function also includes inputs that simultaneously incorporate learning-independent effects of action-specific bias and hysteresis (Colas et al., 2017). For any interactive environment, including these terms is imperative—not only to account for additional variance but also to dissociate illusory mimicry of learning via sequential dependence from actual learning. That is, as learning promotes consistent repetition of responses within a state, so too can autocorrelational effects of hysteresis producing response repetition or alternation that coincidentally aligns with rotating states. (For example, perseveration offers a more parsimonious explanation for action repetition that could otherwise be attributed to an optimistic confirmation bias (Frank et al., 2004; Sharot, 2011; Sharot et al., 2011; Thorndike, 1932, 1933); in RL terms, the latter could translate to an asymmetry in learning rates favoring positive over negative outcomes (Cazé & van der Meer, 2013; Daw et al., 2002; Frank et al., 2007, 2009; Niv et al., 2012)—but at the cost of susceptibility to overfitting (relative to hysteresis) (Chambon et al., 2020; Gershman, 2016; Katahira, 2015, 2018; Palminteri, 2021; Sugawara & Katahira, 2021).) The baseline hysteresis model includes a dynamic perseveration (or alternation) bias $\beta_t(a)$ (cf. Lau & Glimcher, 2005; Schönberg et al., 2007) as well as a constant lateral bias β_R with the arbitrary convention that rightward is positive. These internal biases complemented the learned external action values to dictate the policy's probabilities for each action 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 (Cohen et al., 2007; Daw et al., 2006; Gershman, 2018; Schulz & Gershman, 2019; Speekenbrink & Konstantinidis, 2015; Sutton & Barto, 1998; Thompson, 1933; Wilson et al., 2014). This equation reduces to a logistic function in the present two-alternative forced-choice task:

$$\pi_t(s_t, a) = P(a_t = a | s_t) = \frac{\exp\left\{\left(Q_t(s_t, a) + \beta_t(a) + \beta_R I_R(a)\right)/\tau\right\}}{\sum_{a'} \exp\left\{\left(Q_t(s_t, a') + \beta_t(a') + \beta_R I_R(a')\right)/\tau\right\}}$$

Modeling action hysteresis in terms of the dynamics of cumulative perseveration or alternation biases first requires an initialization

of $\beta_t(a)$, which is here notated so as not to be confused with the parameter β_0 described later:

$$\forall a: \beta_{t=0}(a) = 0$$

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

$$N_0 = 0$$

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

$$\forall a_t: N_t = N_{t-1} + 1$$

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

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

The exponentially decaying hysteretic bias is determined by its initial magnitude β_0 and inverse decay rate λ_β (for $0 \leq \lambda_\beta \leq 1$). A positive magnitude for such autocorrelation represents a perseveration bias in favor of repeating previous actions, whereas a negative magnitude represents an alternation bias in favor of switching between actions—that is, “antiperseveration”. The decay parameter is notated with the convention adopted for the eligibility trace, such that bases λ and λ_β both correspond to the complement of (i.e., unity minus) the exponential decay rate. The exponential decay of the bias proceeds with each action executed, as described in the following equation that integrates cumulative hysteretic biases:

$$\beta_{t+1}(a) = \sum_{i=0}^{N_t-1} \beta_0 \lambda_\beta^i I_{N_t-i}(a)$$

The indicator function $I_R(a)$ is used for a constant lateral bias with the arbitrary convention that a positive sign for β_R corresponds to a rightward bias while a negative sign corresponds to a leftward bias:

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

The final GRL model presently 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_0 coupled with inverse decay rate λ_β for the exponential decay of the perseveration (or alternation) bias. For a paradigm such as this, the eligibility parameter λ cannot be tuned as a free parameter without a multistep Markov decision process (MDP) including intermediate states. As the time steps are discretized with a single step back per trial here, this element was fixed at $\lambda = 0.5$ by default for predicting dynamics of neural activity. This assignment, which did not substantially impact the results if changed, is also in agreement with previous fitted results (mean

$\lambda = 0.684$) arrived at with a two-step MDP and otherwise comparable methodology (Colas et al., 2017).

4.8 | Computational modeling: Model-based learning

Rather than the implicit model of task structure that emerges from the discriminative generalization of GRL, a cognitive map (i.e., model) could instead be represented explicitly as part of a proper model-based algorithm. The following (cognitive) model-based algorithms track a hierarchical metastate that corresponds to the generalizable structure within each state category (e.g., faces or houses). From this intuition, the possible hypotheses h for the binary metastate can be summarized as “AL/BR” or “AR/BL” for a given category c with complementarity between states and actions (Figure 1c), where “AR/BL” means state s_A rewards the right action a_R while state s_B rewards the left action a_L . Whereas the model-free learner naively initialized at zero, the model-based learner initializes the estimated probabilities of the metastate hypotheses $P_t(h|c)$ at $1/2$ for a uniform prior within each category:

$$\forall (c, \{h|c\}) : P_0(h|c) = \frac{1}{|\{h|c\}|} = \frac{1}{2}$$

For both types of model-based systems that follow, the consistent hypothesis $\hat{h}(s_t, a_t, r_{t+1})$ for an observed state-action-outcome sequence is first inferred according to a binary rule. Yet the trial's consistent hypothesis \hat{h} (“h-hat”) is not necessarily true to the block's actual metastate for the category because of the stochastic nature of the environment. In other words, this initial inference only functions as an intermediate input to either model-based learning process:

$$\hat{h}(s_t, a_t, r_{t+1}) = \begin{cases} (AR, BL), & \begin{cases} s_t = s_A, a_t = a_L, r_{t+1} = 0 \\ s_t = s_A, a_t = a_R, r_{t+1} = 1 \\ s_t = s_B, a_t = a_L, r_{t+1} = 1 \\ s_t = s_B, a_t = a_R, r_{t+1} = 0 \end{cases} \\ (AL, BR), & \begin{cases} s_t = s_A, a_t = a_L, r_{t+1} = 1 \\ s_t = s_A, a_t = a_R, r_{t+1} = 0 \\ s_t = s_B, a_t = a_L, r_{t+1} = 0 \\ s_t = s_B, a_t = a_R, r_{t+1} = 1 \end{cases} \end{cases}$$

Given the two possibilities, the trial's alternative hypothesis $\hat{h}'(s_t, a_t, r_{t+1})$ is represented with the previous convention for prime notation such that “h-hat-prime” is complementary to “h-hat”:

$$\hat{h}'(s_t, a_t, r_{t+1}) = \begin{cases} (AL, BR), & \hat{h}(s_t, a_t, r_{t+1}) = (AR, BL) \\ (AR, BL), & \hat{h}(s_t, a_t, r_{t+1}) = (AL, BR) \end{cases}$$

4.9 | Computational modeling: State-prediction error

The simpler model-based algorithm operates with a heuristic analogous to the delta learning of model-free RL but computes a state-prediction error (SPE) δ_t^{SPE} rather than a reward-prediction error (RPE). The SPE is essentially a generalized prediction error in its own right. Whereas for a multistep MDP the transition function for states and state-action pairs (Bellman, 1957; Sutton & Barto, 1998) could be learned with another type of SPE as part of a dynamic-programming algorithm (cf. Gläscher et al., 2010; Lee et al., 2014), this novel type of SPE is instead concerned with the generalizable metastate of the active state's category c_t . Hence the SPE here is a “metastate-prediction error” (MPE). Unlike the signed RPE, the unsigned SPE or MPE takes the difference between unity and the probability estimate for the state-action-outcome sequence's consistent hypothesis:

$$\delta_t^{\text{SPE}} = 1 - P_t(\hat{h}(s_t, a_t, r_{t+1}) | c_t)$$

The update of the probability estimate is weighted by a model-based learning rate α_{SPE} (for $0 \leq \alpha_{\text{SPE}} \leq 1$), which is again analogous to RL:

$$P_{t+1}(\hat{h}(s_t, a_t, r_{t+1}) | c_t) = P_t(\hat{h}(s_t, a_t, r_{t+1}) | c_t) + \alpha_{\text{SPE}} \delta_t^{\text{SPE}}$$

Moreover, the probability estimate for the trial's alternative hypothesis $\hat{h}'(s_t, a_t, r_{t+1})$ is proportionally decreased as well, thus fixing the sum of the probabilities to unity:

$$P_{t+1}(\hat{h}'(s_t, a_t, r_{t+1}) | c_t) = P_t(\hat{h}'(s_t, a_t, r_{t+1}) | c_t) - \alpha_{\text{SPE}} P_t(\hat{h}(s_t, a_t, r_{t+1}) | c_t)$$

4.10 | Computational modeling: Hidden Markov model

The more complex model-based algorithm utilizes Bayesian optimization while specifying an even more explicit and complete model of the exploitable structure in this environment. Whereas previous implementations of the hidden Markov model (HMM) (Ghahramani, 2001) have emphasized reversals between linked states or actions (as a Markov process) (cf. Aquino et al., 2020; Hampton et al., 2006; Prévost et al., 2013), the hidden state in this HMM uniquely corresponds to the hierarchical metastate of a category subsuming active states—that is, a “hidden metastate”. The likelihood function for an outcome given the preceding state-action pair and an assumed hypothesis $P(r_{t+1}|h, (s_t, a_t))$ is determined by a consistency parameter θ_0 (for $\frac{1}{2} \leq \theta_0 \leq 1$) for a binary distribution, representing the agent's belief about the consistency of the rule for a given metastate's probabilistic outcomes:

$$P^{\text{Likelihood}}(r_{t+1} | h, (s_t, a_t)) = \begin{cases} \theta_0, & h = \hat{h}(s_t, a_t, r_{t+1}) \\ 1 - \theta_0, & h = \hat{h}'(s_t, a_t, r_{t+1}) \end{cases}$$

An optimal Bayesian learner tasked with reversal learning such as this can employ belief propagation (Jordan, 1998), such that

knowledge of reversing contingencies is directly factored into the integration of changing evidence. For the full HMM, a second fitted parameter represented the reversal rate θ_1 (for $0 \leq \theta_1 \leq 1$) applied to either complementary hypothesis h' , but the reduced HMM0 variant omits this parameter ($\theta_1 = 0$) so as to not represent any specific expectation of metastate reversals. By the onset of a new trial, the preceding posterior forms the new prior with an update determined by this baseline reversal rate:

$$P_t^{\text{Prior}}(h | c_t) = \theta_1 P_{t-1}^{\text{Posterior}}(h' | c_t) + (1 - \theta_1) P_{t-1}^{\text{Posterior}}(h | c_t)$$

Following Bayes' rule, the updated posterior upon experiencing a new state-action-outcome sequence integrates prior knowledge with the likelihood of the observation given a hypothesis for the category's metastate:

$$P_{t+1}^{\text{Posterior}}(h | c_t) = \frac{P^{\text{Likelihood}}(r_{t+1} | h, (s_t, a_t)) P_t^{\text{Prior}}(h | c_t)}{\sum_{h^*} P^{\text{Likelihood}}(r_{t+1} | h^*, (s_t, a_t)) P_t^{\text{Prior}}(h^* | c_t)}$$

4.11 | Computational modeling: Model-based value

Unlike RL and GRL, the model-based algorithms just described do not learn about value per se; rather, these algorithms track the probabilities of hypothesized metastates and with inference translate these to subjective value in an additional layer of computation. As an alternative to a cached value estimate, the model-based action value $Q_t^{\text{MB}}(c, s, a)$ is inferred for the hierarchy of a category-state-action triplet from probability estimates for the category's metastate. Note that, as an input to model-based decision making in this setting, this action-value estimate is not equivalent to the action's expected value for the actual reward yield of the outcome $E_t[r_{t+1} | c_t, s_t, a]$. For the Bayesian HMM, that expectation also factors in the likelihood function with beliefs about rule consistency:

$$E_t[r_{t+1} | c_t, s_t, a] = \sum_h \sum_r P^{\text{Likelihood}}(r | h, (s_t, a)) P_t^{\text{Prior}}(h | c_t) r$$

Although analogous to the cached reward prediction of RL, the preceding expected value would instead be computed on the fly by the HMM. However, this expectation was not the relevant input to the action-selection policy. Rather, the HMM follows the SPE in more efficiently choosing according to the estimated probability that an action's congruent hypothesis for the category and state is correct. This feature is optimal for the HMM here and was not just implemented in the interest of control in model comparison: For the complementary hypotheses of this paradigm, it proportionately amplifies the difference between action values so as to create greater opportunity for greedy exploitation of presumed knowledge while still achieving exploration through counterfactual learning. Hence the model-based action value is yoked to the dynamic beliefs of either algorithm with a

shared equation more directly translating the probability estimates for the metastate hypotheses:

$$Q_t^{\text{MB}}(c, s, a) = \sum_r P_t(\hat{h}(s, a, r) | c) r = \begin{cases} P_t((AL, BR) | c), & s = s_A, a = a_L \\ P_t((AR, BL) | c), & s = s_A, a = a_R \\ P_t((AR, BL) | c), & s = s_B, a = a_L \\ P_t((AL, BR) | c), & s = s_B, a = a_R \end{cases}$$

4.12 | Computational modeling: Dual systems

The different model-based models and basic RL were all nested within dual-systems models that combine model-based and model-free techniques in parallel. The action-selection policy can thus be expanded with another parameter as the model-based weight w_{MB} (for $0 \leq w_{\text{MB}} \leq 1$), which modulates the weight of either model-based system's estimate of action value. Model-based weighting can reflect the fidelity of the model-based system as well as the arbitrating agent's confidence in its reliability (Daw et al., 2011; Gläscher et al., 2010; Lee et al., 2014). The nested cases of $w_{\text{MB}} = 0$ and $w_{\text{MB}} = 1$ correspond to purely model-free and purely model-based agents, respectively. Whereas a model-free system—including GRL despite linked representations—caches value for only state-action pairs, inferential model-based value explicitly factors in the hierarchical metastates characterizing category-state-action triplets:

$$\pi_t(c_t, s_t, a) = \frac{\exp\left\{\left(w_{\text{MB}} Q_t^{\text{MB}}(c_t, s_t, a) + (1 - w_{\text{MB}}) Q_t(s_t, a) + \beta_t(a) + \beta_{\text{RL}}(a)\right) / \tau\right\}}{\sum_{a^*} \exp\left\{\left(w_{\text{MB}} Q_t^{\text{MB}}(c_t, s_t, a^*) + (1 - w_{\text{MB}}) Q_t(s_t, a^*) + \beta_t(a^*) + \beta_{\text{RL}}(a^*)\right) / \tau\right\}}$$

4.13 | Model fitting

A total of 17 learning models were tested against each other and the hysteresis model ($\alpha = g_A = g_S = 0$) (Table 2). A subset of 11 models corresponded to a factorial model comparison including every nested permutation with respect to the two dimensions of generalization in GRL—to wit, no generalization ($g_A = g_S = 0$), maximally optimal discriminative action generalization ($g_A = -1, g_S = 0$), free action generalization ($-1 \leq g_A \leq 0, g_S = 0$), maximally suboptimal associative state generalization ($g_A = 0, g_S = 1$), maximally optimal discriminative state generalization ($g_A = 0, g_S = -1$), free state generalization ($g_A = 0, -1 \leq g_S \leq 1$), maximally optimal discriminative action generalization and maximally suboptimal associative state generalization ($g_A = -1, g_S = 1$), maximally optimal discriminative action and state generalization ($g_A = g_S = -1$), free action and state generalization with a shared parameter ($g_A = \min\{0, g_S\}, -1 \leq g_S \leq 1$), free action generalization plus free state generalization with dual parameters ($-1 \leq g_A \leq$

0, $-1 \leq g_S \leq 1$), and free interaction between state and action generalization ($g_{SA} \neq g_A$). Competing degenerate models thus benefited from having fewer degrees of freedom to penalize.

The (cognitive) model-based models were tested similarly alongside the others. The SPE model was nested within a dual-systems “SPE+RL” model. Nested within the full HMM was the HMM0 variant without explicit reversals of the hidden state (i.e., metastate) ($\theta_1 = 0$). These two models were nested within their respective dual-systems models that included RL in parallel—that is, “HMM0+RL” and “HMM+RL”. The instantiation of the GRL model optimally tuned with $g_A = g_S = -1$ is also of special note for serving as a model-free approximation (cf. Gläscher et al., 2009; Hampton et al., 2007) of a model-based scheme for an idealized optimal agent in an environment with perfectly anticorrelated states and actions. (Under different circumstances elsewhere, differing degrees of generalization in proportion to the statistics of another environment could be better suited to partially or dynamically correlated or anticorrelated states and actions.) As the most interpretable model comparison is one grounded in a factorial design with systematic testing of parameters, emphasis is due for commensurable models within a single class such as RL (including GRL in this context).

In capturing action-specific bias and hysteresis, the 4-parameter hysteresis model offers a nested null model that is more viable as a control than a zero-parameter chance model with random choices or even an intercept model, which has only one parameter for the probability of an arbitrary action $P(A_1)$. Thus, sensitivity to learnable outcomes or lack thereof can be detected with greater precision by setting the fitting performance of the hysteresis model as a benchmark for comparison with candidate models that feature relevant learning; a participant could then be set aside in the Nonlearner group for demonstrating a lack of reward sensitivity across all learning models. (In the absence of learning, inclusion of τ is redundant in practice but nevertheless maintained as a degree of freedom because of its conceptual relevance as the stochasticity parameter as opposed to a learning parameter per se.)

This uniquely comprehensive modeling approach—that is, the foundation of a 5-parameter model (Colas et al., 2017) rather than the standard 2-parameter model with only learning rate and temperature—also aims to enhance parameter identifiability with respect to actual learning as opposed to other sources of variance that may obscure or mimic learning (Lau & Glimcher, 2005; Schönberg et al., 2007). Whereas alternative solutions find recourse in regularization via fully group-level estimation (i.e., concatenating data sets or averaging parameters) or the intermediate approach of hierarchical Bayesian modeling across individuals (Ahn et al., 2017; Daw, 2011; Gershman, 2016), the present solution of a more complete yet parsimonious model—in this case accounting for action-specific bias and hysteresis—avoids compromising the independence of separate data sets. As per the bias-variance tradeoff, even reducing variance with the constraints of hierarchical group-level estimation would necessarily introduce bias both toward the average across individuals and toward the specifications of a parametric probability distribution. In

other words, the present technique cannot be diminished by potentially inappropriate assumptions that a given participant is learning and furthermore learning in a particular way merely because other participants in the aggregate have mostly demonstrated learning and an overall tendency to learn in a particular way. Added complexities such as idiosyncratic strategies for generalization impose even greater demands for accommodating individual differences. This subject-level interpretability also extends to (computational) model-based analysis of neurophysiological data (O'Doherty et al., 2007), where advantages can include more precise estimation of signal dynamics and parameters of interest—including between-subject analyses—as well as the capacity to classify distinct types of performance in subgroup analyses—for example, learners versus nonlearners (Colas et al., 2017) or associative generalizers versus discriminative generalizers.

The competing models were all fitted to empirical behavior at the level of individual subjects via maximum-likelihood estimation. Free parameters were optimized for overall goodness of fit to a subject's sequence of actions with randomly seeded iterations of the Nelder-Mead simplex algorithm (Nelder & Mead, 1965). All modeling and fitting procedures were programmed with MATLAB. The Akaike information criterion with correction for finite sample size (AICc) (Akaike, 1974; Hurvich & Tsai, 1989) provided a means to adjust for model complexity when comparing models that differ in degrees of freedom. The preferred model was also to provide the basis for the subsequent neuroimaging analysis.

To verify the discriminability of the preferred 7-parameter GRL model, each fitted instantiation of the model was subsequently used to simulate a data set yoked to that of the respective subject. Another complete model comparison was conducted for these simulated data as a test of model recovery that would indicate whether this model could be discriminated reliably among the competing alternatives. The same procedure was repeated with simulations conversely derived from 5-parameter basic RL for additional reassurance that the original results could not be reduced to mere overfitting.

4.14 | Data analysis: Behavior

Performance on the learning task was assessed for each participant by calculating overall accuracy as the proportion of choices of 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 the chance level (Schönberg et al., 2007); subsequent modeling could also confirm that this label was appropriate for each individual within the group. The remaining participants with accuracy not significantly greater than chance were subsequently assigned to either the “Poor learner” group or the “Nonlearner” group according to whether or not a learning model could yield a significant improvement in goodness of fit relative to a hysteresis model without sensitivity to actual

outcomes (Colas et al., 2017). Reaction time (RT) was also compared between these primary groups via one-tailed independent-samples *t* tests hypothesizing a speed-accuracy tradeoff.

Across the two learner groups, the second stage of model comparison reclassified these individuals in secondary “Discriminative generalizer”, “Nongeneralizer”, and “Associative generalizer” groups for the cases of $g_S < 0$, $g_S = 0$, and $g_S > 0$, respectively, as determined by the individually fitted GRL model. A subset of diagnostic trials were selected to represent the first opportunities for generalization of reward within each block. Sixteen trials in total corresponded to the two categories each having two newly rewarded actions per each of four runs (i.e., $2 \times 2 \times 4$). First-generalization accuracy was compared against chance with one-tailed one-sample *t* tests within each model-defined group; Discriminative generalizers were hypothesized to perform above chance, whereas Nongeneralizers and especially Associative generalizers were hypothesized to perform below chance in the absence of direct reinforcement for the new reward contingencies. One-tailed independent-samples *t* tests followed to verify the presumed ranking of Discriminative generalizers, Nongeneralizers, then Associative generalizers. Moreover, a correlation between the GRL model's fitted parameter and first-generalization accuracy was tested for using linear regression with a one-tailed one-sample *t* test and the Pearson correlation coefficient. For a posterior predictive check of each generative model with respect to these results, simulated data sets were yoked to the empirical data sets and analyzed in the same fashion after averaging across 1,000 simulations.

Taking the free parameters fitted for each subject, the overall reward sensitivity of each instantiation of the GRL model was quantified as $\log(\alpha(1-g_A-g_S+g_Sg_A)/\tau)$ (cf. Colas et al., 2017; Schönberg et al., 2007) with a logarithmic transformation for more interpretable rescaling prior to presentation of the results. Relative to the softmax temperature τ , this formula factors in the magnitudes of all four possible updates of action values with each duplicated and relayed RPE signal—that is, for the current (α) and complementary ($g_A\alpha$) actions within the current state as well as the current ($g_S\alpha$) and complementary ($g_Sg_A\alpha$) actions within the complementary state. Considering that fitted RL models are typically characterized by a correlation between learning rate and softmax temperature that reflects elongated maxima in their joint likelihood function (Daw, 2011), this sensitivity ratio is a more precise and more relevant measure of a learning model's sensitivity than either the learning rate or the temperature alone. Such an alpha-tau correlation was observed across Learner groups in both data sets (3FH: $r = 0.518$, $t_{38} = 3.73$, $p < 10^{-3}$; 7CM: $r = 0.427$, $t_{19} = 2.05$, $p = .027$). Accordingly, sensitivity was compared between the Good-learner and Poor-learner groups by way of a one-tailed independent-samples *t* test. Post-hoc one-tailed independent-samples *t* tests were subsequently conducted for action generalization and state generalization. To test for correlations between sensitivity and accuracy or RT, between action generalization and accuracy or RT, between state generalization and accuracy or RT, and between action generalization and state generalization, linear regression was performed with one-tailed one-sample *t* tests and reported with the Pearson correlation coefficient.

Across all trials, analyses of choice data based on preceding outcomes first separated the most recent trials in which either the same (i.e., current) state was encountered or the other, complementary state within the current category was encountered. These trials were further binned according to whether the trial rewarded a given action or provided no such reward. The probability of repeating the prior trial's action was calculated within each of four bins: “same/reward”, “same/no-reward”, “other/reward”, and “other/no-reward”. Given the complementarity of states within a category to facilitate discriminative state generalization by design, the hypothesis for repeating actions from the “other” state was an inversion of the hypothesis for the “same” state: A previous reward in the same state was supposed to increase repetition of the action, whereas a previous reward in the other state was supposed to decrease repetition. For each participant group, one-tailed one-sample *t* tests compared the probability of repeating the respective state's last action between the “reward” and “no-reward” conditions either within same-state trials or within other-state trials. Moreover, another set of one-tailed one-sample *t* tests assessed the between-state interaction of the effect between the “reward” and “no-reward” conditions. To verify that the GRL model could quantitatively reproduce these results as well, simulated data sets yoked to the empirical data sets were analyzed in the same manner for a second posterior predictive check.

As computational modeling provided quantitative trial-by-trial estimates of action-value representations, these dynamic variables could in turn be related to psychometric functions for choices and RTs. A logistic-regression model first modeled the probability of repeating the most recent action (independent of state) as a function of the normalized difference between action values. A linear-regression model likewise modeled the RT as a function of the normalized absolute value of the difference between action values. In order to accommodate intersubject variability in the range of estimated values, differences in action values were normalized with respect to the maximum absolute value for each subject. Parameters for these mixed-effects models were first estimated at the level of individual subjects and subsequently assessed within each subject group using one-tailed one-sample *t* tests.

4.15 | Data analysis: Neuroimaging

Analysis of the fMRI data was conducted with SPM and carried out identically within each of the 3-T and 7-T data sets. This (computational) model-based analysis (O'Doherty et al., 2007) was grounded in the explicit quantitative dynamics predicted by the GRL model with subject-specific parameters (Figure 7). The GLM of BOLD signals was essentially a tripartite model characterized by parametric regressors for value, RPE, and RT. For a paradigm such as this with a single-step cue-outcome sequence, disambiguating all three types of signals is nontrivial (as alluded to previously with reference to the CQ model).

Indicator variables modeled as boxcar functions described all of the events within the sequence of each trial. These indicators included decision time (with variable duration), face (or color) cues (with a

duration of 2 s), house (or motion) cues (2 s), left-hand responses (2 s), right-hand responses (2 s), the ISI (3 s), outcomes (1 s), and the ITI (3–7 s). In the case of a missed trial marked by failure to respond in the 2-s window, events were coded as separate indicator variables for face (or color) cues (2 s), house (or motion) cues (2 s), error feedback (1 s), late left-hand responses within a 1-s window (1 s), late right-hand responses within a 1-s window (1 s), and the ITI (6–10 s). In preventing the complications of temporal prediction-error signals such as in TDRL (McClure et al., 2003; O'Doherty et al., 2003; O'Doherty et al., 2004; Sutton, 1988; Sutton & Barto, 1998), the fixed ISI was sufficient and did not result in rank deficiency for the design matrix because of not only jitter in the ITI but also the quantitative precision of narrowly specified (computational) model-based regressors.

The RT regressor was specified as a boxcar function aligned with cue onset and extending with a variable duration corresponding to trial-by-trial RT. Value signals were continuous and included the state value $V_t(s_0)$ of the preparatory state, the chosen action value $Q_t(s_t, a_t)$ for the active state, and the value of the outcome state. Similarly, learning signals in the form of RPE signals included both the SVPE δ_t^V computed upon encountering the cue—as per the TD algorithm—and the AVPE δ_t^Q computed upon encountering the outcome. Value and learning signals were modeled as parametric modulators of boxcar functions, and the duration of each boxcar function corresponded to the duration of the respective stimulus with one exception: Value signals were assumed to persist beyond stimulus offset through the subsequent ISI. The reason for this convention is that the expectation for value should remain the same with negligible temporal discounting. Although the distinctions between state value and action value—or between the SVPE and the AVPE—are important in general (Averbeck & O'Doherty, 2022; Colas et al., 2017; O'Doherty et al., 2004), such distinctions are beyond the scope of the present study and were necessarily omitted here in consideration of the single-step cue-outcome sequence, which challenges dissociability.

As orthogonalization was forgone to avoid potential distortions of the parameter estimates or their interpretation (Mumford et al., 2015), the complete predictions of this TDRL model were taken advantage of to minimize inevitable multicollinearity (Colas et al., 2017; cf. Behrens et al., 2008; Zhang et al., 2020). In addition to effects of value on RT (Busemeyer & Townsend, 1993; Colas, 2017; Laming, 1968; Luce, 1986; Ratcliff, 1978; Usher & McClelland, 2001), there is also a relation between value and the RPE; the latter is a linear combination (i.e., subtraction) of outcome value and estimated value. By collapsing events across a trial into unitary regressors for value and learning signals, the correlation between value and the RPE could be mitigated to a tractable level of dissociability (3FH: mean $r^2 = 0.431$ across subjects; 7CM: mean $r^2 = 0.389$). Dissociation was also achieved between value and RT (3FH: mean $r^2 = 0.008$; 7CM: mean $r^2 = 0.007$) as well as between the RPE and RT (3FH: mean $r^2 = 0.014$; 7CM: mean $r^2 = 0.021$), such that triply dissociated value, decision, and learning signals could all be accounted for in parallel.

All of the aforementioned predictor variables were convolved with a canonical double-gamma hemodynamic-response function as inputs to the GLM. The design matrix also included the confound

regressors without convolution—to wit, motion parameters, CSF signal, white-matter signal, global signal, framewise displacement, standardized DVARS, indicators for nonsteady states (i.e., outlier volumes), DCT basis functions, tCompCor components, aCompCor components, ICA-AROMA noise components, and RETROICOR components from cardiac and respiratory data. Also among the nonconvolved regressors were a first-degree autoregressive (i.e., “AR(1)”) term and a constant term. GLMs were first estimated at the level of an individual subject, and contrasts of parameter estimates were subsequently computed for the parametric regressors at the group level as part of a mixed-effects analysis. The groups corresponded to Good learners, Poor learners, Nonlearners, or all learners (including both Good and Poor learners). Positive effects of these contrasts were tested for using one-tailed one-sample *t* tests.

Strictly aiming for the rigor of quantitative parametric regressors, the default thresholds for statistical significance and cluster extent were preset at standard levels of $p < .005$ and $k \geq 10$ voxels (Forman et al., 1995; Lieberman & Cunningham, 2009). Whereas whole-brain correction for multiple comparisons was precluded by so many voxels being sampled with high resolution—and especially so at 7 T—precise regions of interest (ROIs) could constrain the hypothesis space a priori with established precedents for the neural correlates of evaluation and value-based decision making and learning. Spherical coordinate-based ROIs with 6-mm radii were applied to small-volume correction (SVC) controlling for the familywise error rate (FWE) at $p < .05$. A given set of ROIs was first tested as a network; post-hoc tests followed for individual ROIs within the set. For visualization, statistical-parametric maps were overlaid on averages of processed anatomical images from the respective participants included in a given analysis.

For learning signals, a prior high-resolution study with comparable methodology (Colas et al., 2017) provided focal coordinates for variants of the RPE signal in the left anterior caudate nucleus at (−8, 18, −8) or (−8.4, 18, −8.4) for 3-T or 7-T images, respectively; the right nucleus accumbens at (8, 12, −4) or (8.4, 12, −3.6); the right ventral putamen at (18, 12, −12); the left nucleus accumbens at (−12, 10, −6) or (−12, 9.6, −6); the right dorsal putamen at (28, 6, 0) or (27.6, 6, 0); the left dorsal caudate nucleus at (−18, 2, 16) or (−18, 2.4, 15.6); and the left SN at (−10, −14, −12) or (−9.6, −14.4, −12). More broadly, exploratory anatomical ROIs to be searched for uncorrected results included the entire striatum and the dopaminergic mid-brain, comprising the SN and the ventral tegmental area.

For value signals, a pair of meta-analytic studies with largely compatible results (Bartra et al., 2013; Clithero & Rangel, 2014) provided coordinates for the correlates of monetary value in bilateral ventromedial prefrontal cortex (vmPFC) at (0, 46, −8) or (0, 45.6, −8.4) for 3-T or 7-T images, respectively; the right nucleus accumbens at (10, 16, −6) or (9.6, 15.6, −6); the left nucleus accumbens at (−10, 10, −6) or (−9.6, 9.6, −6); and bilateral posterior cingulate cortex (PCC) at (−2, −34, 38) or (−2.4, −33.6, 38.4). Coordinates were averaged between the two meta-analyses, which identified a common set of regions for the pertinent contrast. Exploratory ROIs to be searched for uncorrected results included the entirety of vmPFC, the striatum, and PCC.

For decision signals, a meta-analysis (Yarkoni et al., 2009) provided coordinates for the correlates of RT across different tasks in bilateral medial frontal cortex (MFC) at (0, 12, 48). Encompassing the vicinity of the supplementary motor area (SMA), the pre-SMA, and dorsal anterior cingulate cortex, MFC as a whole also served as an exploratory ROI to be searched for uncorrected results. Although other brain areas such as premotor cortex and posterior parietal cortex have been implicated in decision-making processes as well (Cisek, 2012; Cisek & Kalaska, 2010; Gold & Shadlen, 2007), greater effector-specific lateralization in these regions limits their interpretability with respect to the more abstract value-based decision making sought here. In any case, the scope of the present study is limited such that the gamut of diverse decision-making signals is not investigated in the fullest detail.

Regressors for specific effects of state generalization and action generalization were quantified as $-g_S/\tau$ and $-g_A/\tau$, respectively, to test for interactions with RPE signals at the second level between subjects. The ROIs applied to the original RPE contrast were utilized here as well. However, the paradigm of the study that these ROIs were derived from (Colas et al., 2017) did not include the present factor of generalization in any form. As such, the hypotheses motivating these ROIs in their original context were less definitive for exploration in this new context. On the other hand, the ROIs still can serve as candidates for first-pass investigation. Lacking proper precedent, exploratory investigation throughout the striatum and especially the dopaminergic midbrain was considered more openly here.

Additionally for effects of generalization, the most active locus within the hippocampus was extracted from broad meta-analytic results in the Neurosynth database (Yarkoni et al., 2011). Across results including the term “hippocampus” in the report’s abstract, the peak activations derived from uniformity and association tests coincided at (−28, −18, −16), which was reflected across hemispheres with bilateral SVC as (±28, −18, −16) or (±27.6, −18, −15.6) for 3-T or 7-T images, respectively. Another exploratory ROI further included the entire hippocampal region.

With regard to state generalization, presenting categorical stimuli elicited activation in the expected cortical regions: The categories of faces, houses, colors, and directions of motion activated the FFA (Kanwisher et al., 1997), the PPA (Epstein & Kanwisher, 1998), color-sensitive visual area V4 (or V8) (Beauchamp et al., 1999; Hadjikhani et al., 1998; Wade et al., 2002; Zeki et al., 1991), and the motion-sensitive middle-temporal area MT (or V5) (Tootell et al., 1995; Watson et al., 1993; Zeki et al., 1991), respectively ($p < .005$). With regard to action generalization, actions executed with the left and right hands generated the expected activation in contralateral motor cortex and ipsilateral cerebellar cortex (Grafton et al., 1992) at both 3 T and 7 T ($p < .005$). In the interest of being concise, results for these less critical contrasts are not reported in further detail.

As a control analysis, a regressor for the normalized learning rate was quantified as α/τ to test for interactions with RPE signals at the second level. This contrast was juxtaposed with the generalization contrasts for the same sets of ROIs—the goal being to determine if

the findings for generalization specifically were possibly confounded with a nonspecific effect of learning performance.

ACKNOWLEDGMENTS

This study originated at a workshop on “Learning in Networks” supported by the National Institute for Mathematical and Biological Synthesis. 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 MH094258). JMT was supported by National Institute of Mental Health grant P50 MH094258. AWT was supported by National Institute of Biomedical Imaging and Bioengineering grant P41 EB015922. JIG was supported by National Institute of Mental Health grant R01 MH115557. DSB was supported by Army Research Office grants W911NF-16-1-0474 and W911NF-18-1-0244. CAH was supported by the Klingenstein-Simons Neuroscience Fellowship.

DATA AVAILABILITY STATEMENT

Data are available at <https://neurovault.org/collections/RLVWMYCQ/>.

ORCID

Jaron T. Colas  <https://orcid.org/0000-0003-1872-7614>

Neil M. Dundon  <https://orcid.org/0000-0001-6246-1775>

Raphael T. Gerraty  <https://orcid.org/0000-0001-9782-1005>

Natalie M. Saragosa-Harris  <https://orcid.org/0000-0002-4493-6113>

Karol P. Szymula  <https://orcid.org/0000-0003-1822-0688>

Koranis Tanwisuth  <https://orcid.org/0000-0003-3563-6781>

J. Michael Tyszka  <https://orcid.org/0000-0001-9342-9014>

Camilla van Geen  <https://orcid.org/0000-0002-4948-5550>

Harang Ju  <https://orcid.org/0000-0003-1904-1753>

Arthur W. Toga  <https://orcid.org/0000-0001-7902-3755>

Joshua I. Gold  <https://orcid.org/0000-0002-6018-0483>

Dani S. Bassett  <https://orcid.org/0000-0002-6183-4493>

Catherine A. Hartley  <https://orcid.org/0000-0003-0177-7295>

Daphna Shohamy  <https://orcid.org/0000-0003-4239-4960>

Scott T. Grafton  <https://orcid.org/0000-0003-4015-3151>

John P. O’Doherty  <https://orcid.org/0000-0003-0016-3531>

REFERENCES

- Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8, 14. <https://doi.org/10.3389/fninf.2014.00014>
- Ahmed, N., Natarajan, T., & Rao, K. R. (1974). Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1), 90–93. <https://doi.org/10.1109/t-c.1974.223784>
- Ahn, W. Y., Haines, N., & Zhang, L. (2017). Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry*, 1, 24–57. https://doi.org/10.1162/cpsy_a_00002

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/tac.1974.1100705>
- Akam, T., Costa, R., & Dayan, P. (2015). Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLoS Computational Biology*, 11(12), e1004648. <https://doi.org/10.1371/journal.pcbi.1004648>
- Andersson, J. L. R., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. *NeuroImage*, 20(2), 870–888. [https://doi.org/10.1016/s1053-8119\(03\)00336-7](https://doi.org/10.1016/s1053-8119(03)00336-7)
- Aquino, T. G., Minxha, J., Dunne, S., Ross, I. B., Mamelak, A. N., Rutishauser, U., & O'Doherty, J. P. (2020). Value-related neuronal responses in the human amygdala during observational learning. *Journal of Neuroscience*, 40(24), 4761–4772. <https://doi.org/10.1523/jneurosci.2897-19.2020>
- Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1), 26–41. <https://doi.org/10.1016/j.media.2007.06.004>
- Avants, B. B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., & Gee, J. C. (2010). The optimal template effect in hippocampus studies of diseased populations. *NeuroImage*, 49(3), 2457–2466. <https://doi.org/10.1016/j.neuroimage.2009.09.062>
- Averbeck, B., & O'Doherty, J. P. (2022). Reinforcement-learning in frontostriatal circuits. *Neuropsychopharmacology*, 47(1), 147–162. <https://doi.org/10.1038/s41386-021-01108-0>
- Balcarras, M., & Womelsdorf, T. (2016). A flexible mechanism of rule selection enables rapid feature-based reinforcement learning. *Frontiers in Neuroscience*, 10, 125. <https://doi.org/10.3389/fnins.2016.00125>
- Ballard, I. C., & McClure, S. M. (2019). Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *Journal of Neuroscience Methods*, 317, 37–44. <https://doi.org/10.1016/j.jneumeth.2019.01.006>
- Ballard, I. C., Wagner, A. D., & McClure, S. M. (2019). Hippocampal pattern separation supports reinforcement learning. *Nature Communications*, 10(1), 1–12. <https://doi.org/10.1038/s41467-019-08998-1>
- Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M., & Behrens, T. E. J. (2021). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron*, 109(4), 713–723. <https://doi.org/10.1016/j.neuron.2020.11.024>
- Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, 7(4), 404–410. <https://doi.org/10.1038/nn1209>
- Barry, R. L., Coaster, M., Rogers, B. P., Newton, A. T., Moore, J., Anderson, A. W., Zald, D. H., & Gore, J. C. (2013). On the origins of signal variance in fMRI of the human midbrain at high field. *PLoS One*, 8(4), e62708. <https://doi.org/10.1371/journal.pone.0062708>
- Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13(5), 834–846. <https://doi.org/10.1109/tsmc.1983.6313077>
- Barto, A. G., Sutton, R. S., & Anderson, C. W. (2021). Looking back on the actor-critic architecture. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1), 40–50. <https://doi.org/10.1109/tsmc.2020.3041775>
- Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, 76, 412–427. <https://doi.org/10.1016/j.neuroimage.2013.02.063>
- Bassett, D. S., Zurn, P., & Gold, J. I. (2018). On the nature and use of models in network neuroscience. *Nature Reviews Neuroscience*, 19(9), 566–578. <https://doi.org/10.1038/s41583-018-0038-8>
- Beauchamp, M. S., Haxby, J. V., Jennings, J. E., & DeYoe, E. A. (1999). An fMRI version of the Farnsworth-Munsell 100-hue test reveals multiple color-selective areas in human ventral occipitotemporal cortex. *Cerebral Cortex*, 9(3), 257–263. <https://doi.org/10.1093/cercor/9.3.257>
- Beckmann, C. F., & Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2), 137–152. <https://doi.org/10.1109/tmi.2003.822821>
- Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, 456(7219), 245–249. <https://doi.org/10.1038/nature07538>
- Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, 100(2), 490–509. <https://doi.org/10.1016/j.neuron.2018.10.002>
- Behzadi, Y., Restom, K., Liu, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, 37(1), 90–101. <https://doi.org/10.1016/j.neuroimage.2007.04.042>
- Beisteiner, R., Robinson, S., Wurnig, M., Hilbert, M., Merksa, K., Rath, J., Höllinger, I., Klinger, N., Marosi, C., Trattng, S., & Geißler, A. (2011). Clinical fMRI: Evidence for a 7 T benefit over 3 T. *NeuroImage*, 57(3), 1015–1021. <https://doi.org/10.1016/j.neuroimage.2011.05.010>
- Bellman, R. E. (1957). *Dynamic programming*. Princeton University Press.
- Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4), 954–967. <https://doi.org/10.1016/j.cell.2020.09.031>
- Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.
- Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, 22(6), 956–962. <https://doi.org/10.1016/j.conb.2012.05.008>
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4), 433–436. <https://doi.org/10.1163/156856897x00357>
- Bromberg-Martin, E. S., Matsumoto, M., Hong, S., & Hikosaka, O. (2010). A pallidum-habenula-dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology*, 104(2), 1068–1076. <https://doi.org/10.1152/jn.00158.2010>
- Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3), 432–459. <https://doi.org/10.1037/0033-295x.100.3.432>
- Bush, R. R., & Mosteller, F. (1951a). A mathematical model for simple learning. *Psychological Review*, 58(5), 313–323. <https://doi.org/10.1037/h0054388>
- Bush, R. R., & Mosteller, F. (1951b). A model for stimulus generalization and discrimination. *Psychological Review*, 58(6), 413–423. <https://doi.org/10.1037/h0054576>
- Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J. R., & Sirigu, A. (2004). The involvement of the orbitofrontal cortex in the experience of regret. *Science*, 304(5674), 1167–1170. <https://doi.org/10.1126/science.1094550>
- Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1), 51–62. <https://doi.org/10.1038/nrn3136>
- Carp, J., Kim, K., Taylor, S. F., Fitzgerald, K. D., & Weissman, D. H. (2010). Conditional differences in mean reaction time explain effects of response congruency, but not accuracy, on posterior medial frontal cortex activity. *Frontiers in Human Neuroscience*, 4, 231. <https://doi.org/10.3389/fnhum.2010.00231>
- Cazé, R. D., Khamassi, M., Aubin, L., & Girard, B. (2018). Hippocampal replays under the scrutiny of reinforcement learning models. *Journal of Neurophysiology*, 120(6), 2877–2896. <https://doi.org/10.1152/jn.00145.2018>
- Cazé, R. D., & van der Meer, M. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics*, 107(6), 711–719. <https://doi.org/10.1007/s00422-013-0571-5>
- Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., & Palminteri, S. (2020). Information about action outcomes differentially

- affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, 4(10), 1067–1079. <https://doi.org/10.1038/s41562-020-0919-5>
- Charpentier, C. J., Iigaya, K., & O'Doherty, J. P. (2020). A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning. *Neuron*, 106(4), 687–699. <https://doi.org/10.1016/j.neuron.2020.02.028>
- Chase, H. W., Kumar, P., Eickhoff, S. B., & Dombrovski, A. Y. (2015). Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience*, 15(2), 435–459. <https://doi.org/10.3758/s13415-015-0338-7>
- Chien, S., Wiehler, A., Spezio, M., & Gläscher, J. (2016). Congruence of inherent and acquired values facilitates reward-based decision-making. *Journal of Neuroscience*, 36(18), 5003–5012. <https://doi.org/10.1523/jneurosci.3084-15.2016>
- Choi, U. S., Kawaguchi, H., Matsuoka, Y., Kober, T., & Kida, I. (2019). Brain tissue segmentation based on MP2RAGE multi-contrast images in 7 T MRI. *PLoS One*, 14(2), e0210803. <https://doi.org/10.1371/journal.pone.0210803>
- Cisek, P. (2012). Making decisions through a distributed consensus. *Current Opinion in Neurobiology*, 22(6), 927–936. <https://doi.org/10.1016/j.conb.2012.05.007>
- Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, 33, 269–298. <https://doi.org/10.1146/annurev.neuro.051508.135409>
- Clithero, J. A., & Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, 9(9), 1289–1302. <https://doi.org/10.1093/scan/nst106>
- Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481), 933–942. <https://doi.org/10.1098/rstb.2007.2098>
- Colas, J. T. (2017). Value-based decision making via sequential sampling with hierarchical competition and attentional modulation. *PLoS One*, 12(10), e0186822. <https://doi.org/10.1371/journal.pone.0186822>
- Colas, J. T., Pauli, W. M., Larsen, T., Tyszka, J. M., & O'Doherty, J. P. (2017). 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 Computational Biology*, 13(10), e1005810. <https://doi.org/10.1371/journal.pcbi.1005810>
- Colizoli, O., de Gee, J. W., van der Zwaag, W., & Donner, T. H. (2021). Functional magnetic resonance imaging responses during perceptual decision-making at 3 and 7 T in human cortex, striatum, and brainstem. *Human Brain Mapping*, 43(4), 1265–1279. <https://doi.org/10.1002/hbm.25719>
- Collette, S., Pauli, W. M., Bossaerts, P., & O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *eLife*, 6, e29718. <https://doi.org/10.7554/elife.29718>
- Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, 21(10), 576–586. <https://doi.org/10.1038/s41583-020-0355-6>
- Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, 120(1), 190–229. <https://doi.org/10.1037/a0030852>
- Collins, D. L., Neelin, P., Peters, T. M., & Evans, A. C. (1994). Automatic 3D intersubject registration of MR volumetric data in standardized talairach space. *Journal of Computer Assisted Tomography*, 18(2), 192–205. <https://doi.org/10.1097/00004728-199403000-00005>
- Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: A neuroimaging study of choice behavior. *Nature Neuroscience*, 8(9), 1255–1262. <https://doi.org/10.1038/nn1514>
- Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3), 162–173. <https://doi.org/10.1006/cbmr.1996.0014>
- D'Ardenne, K., Lohrenz, T., Bartley, K. A., & Montague, P. R. (2013). Computational heterogeneity in the human mesencephalic dopamine system. *Cognitive, Affective, & Behavioral Neuroscience*, 13(4), 747–756. <https://doi.org/10.3758/s13415-013-0191-5>
- Da Costa, S., Saenz, M., Clarke, S., & van der Zwaag, W. (2015). Tonotopic gradients in human primary auditory cortex: Concurring evidence from high-resolution 7 T and 3 T fMRI. *Brain Topography*, 28(1), 66–69. <https://doi.org/10.1007/s10548-014-0388-0>
- da Silva, C. F., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, 4(10), 1053–1066. <https://doi.org/10.1038/s41562-020-0905-y>
- Dagli, M. S., Ingelholm, J. E., & Haxby, J. V. (1999). Localization of cardiac-induced signal change in fMRI. *NeuroImage*, 9(4), 407–415. <https://doi.org/10.1006/nimg.1998.0424>
- Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, 9(2), 179–194. <https://doi.org/10.1006/nimg.1998.0395>
- Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning: Attention and performance XXIII* (pp. 3–38). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199600434.001.0001>
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6), 1204–1215. <https://doi.org/10.1016/j.neuron.2011.02.027>
- Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, 15(4–6), 603–616. [https://doi.org/10.1016/s0893-6080\(02\)00052-7](https://doi.org/10.1016/s0893-6080(02)00052-7)
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12), 1704–1711. <https://doi.org/10.1038/nn1560>
- Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095), 876–879. <https://doi.org/10.1038/nature04766>
- Daw, N. D., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, 26(5), 593–620. <https://doi.org/10.1521/soco.2008.26.5.593>
- Dayan, P. (1992). The convergence of TD(λ) for general λ . *Machine Learning*, 8(3–4), 341–362. <https://doi.org/10.1023/a:1022632907294>
- Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4), 613–624. <https://doi.org/10.1162/neco.1993.5.4.613>
- Dayan, P., & Sejnowski, T. J. (1994). TD(λ) converges with probability 1. *Machine Learning*, 14(3), 295–301. <https://doi.org/10.1023/a:1022657612745>
- de Hollander, G., Keuken, M. C., & Forstmann, B. U. (2015). The subcortical cocktail problem; mixed signals from the subthalamic nucleus and substantia nigra. *PLoS One*, 10(3), e0120572. <https://doi.org/10.1371/journal.pone.0120572>
- de Hollander, G., Keuken, M. C., van der Zwaag, W., Forstmann, B. U., & Trampel, R. (2017). Comparing functional MRI protocols for small, iron-rich basal ganglia nuclei such as the subthalamic nucleus at 7 T and 3 T. *Human Brain Mapping*, 38(6), 3226–3248. <https://doi.org/10.1002/hbm.23586>
- De Martino, F., Yacoub, E., Kemper, V., Moerel, M., Uludağ, K., De Weerd, P., Uğurbil, K., Goebel, R., & Formisano, E. (2018). The impact of ultra-high field MRI on cognitive and computational neuroimaging. *NeuroImage*, 168, 366–382. <https://doi.org/10.1016/j.neuroimage.2017.03.060>

- Delgado, M. R., & Dickerson, K. C. (2012). Reward-related learning via multiple memory systems. *Biological Psychiatry*, 72(2), 134–141. <https://doi.org/10.1016/j.biopsych.2012.01.023>
- Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, 18(5), 767–772. <https://doi.org/10.1038/nn.3981>
- Doll, B. B., Shohamy, D., & Daw, N. D. (2015). Multiple memory systems as substrates for multiple decision systems. *Neurobiology of Learning and Memory*, 117, 4–13. <https://doi.org/10.1016/j.nlm.2014.04.014>
- Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22(6), 1075–1081. <https://doi.org/10.1016/j.conb.2012.08.003>
- Dumoulin, S. O., Fracasso, A., van der Zwaag, W., Siero, J. C., & Petridou, N. (2018). Ultra-high field MRI: Advancing systems neuroscience towards mesoscopic human brain function. *NeuroImage*, 168, 345–357. <https://doi.org/10.1016/j.neuroimage.2017.01.028>
- Düzel, E., Bunzeck, N., Guitart-Masip, M., Wittmann, B., Schott, B. H., & Tobler, P. N. (2009). Functional imaging of the human dopaminergic midbrain. *Trends in Neurosciences*, 32(6), 321–328. <https://doi.org/10.1016/j.tins.2009.02.005>
- Düzel, E., Guitart-Masip, M., Maass, A., Hämmerer, D., Betts, M. J., Speck, O., Weiskopf, N., & Kanowski, M. (2015). Midbrain fMRI: Applications, limitations and challenges. In K. Uludağ, K. Uğurbil, & L. Berliner (Eds.), *fMRI: From nuclear spins to brain functions* (pp. 581–609). Springer. <https://doi.org/10.1007/978-1-4899-7591-1>
- Eapen, M., Zald, D. H., Gatenby, J. C., Ding, Z., & Gore, J. C. (2011). Using high-resolution MR imaging at 7T to evaluate the anatomy of the mid-brain dopaminergic system. *American Journal of Neuroradiology*, 32(4), 688–694. <https://doi.org/10.3174/ajnr.a2355>
- Eckstein, M. K., & Collins, A. G. (2020). Computational evidence for hierarchically structured reinforcement learning in humans. *Proceedings of the National Academy of Sciences*, 117(47), 29381–29389. <https://doi.org/10.1073/pnas.1912330117>
- Eldar, E., Lièvre, G., Dayan, P., & Dolan, R. J. (2020). The roles of online and offline replay in planning. *eLife*, 9, e56911. <https://doi.org/10.7554/eLife.56911>
- Enzmann, D. R., & Pelc, N. J. (1992). Brain motion: Measurement with phase-contrast MR imaging. *Radiology*, 185(3), 653–660. <https://doi.org/10.1148/radiology.185.3.1438741>
- Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598–601. <https://doi.org/10.1038/33402>
- Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1), 111–116. <https://doi.org/10.1038/s41592-018-0235-4>
- Feinberg, D. A., & Setsompop, K. (2013). Ultra-fast MRI of the human brain with simultaneous multi-slice imaging. *Journal of Magnetic Resonance*, 229, 90–100. <https://doi.org/10.1016/j.jmr.2013.02.002>
- Filliter, J. H., Glover, J. M., McMullen, P. A., Salmon, J. P., & Johnson, S. A. (2016). The DalHouses: 100 new photographs of houses with ratings of typicality, familiarity, and degree of similarity to faces. *Behavior Research Methods*, 48(1), 178–183. <https://doi.org/10.3758/s13428-015-0561-8>
- Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2), 774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>
- Fonov, V. S., Evans, A. C., McKinstry, R. C., Almlí, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47(Supplement 1), S102. [https://doi.org/10.1016/s1053-8119\(09\)70884-5](https://doi.org/10.1016/s1053-8119(09)70884-5)
- Fontanesi, L., Gluth, S., Rieskamp, J., & Forstmann, B. U. (2019). The role of dopaminergic nuclei in predicting and experiencing gains and losses: A 7T human fMRI study. *bioRxiv*, 732560. <https://doi.org/10.1101/732560>
- Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, 26(4), 1099–1121. <https://doi.org/10.3758/s13423-018-1554-2>
- Fontanesi, L., Palminteri, S., & Lebreton, M. (2019). Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: A meta-analytical approach using diffusion decision modeling. *Cognitive, Affective, & Behavioral Neuroscience*, 19(3), 490–502. <https://doi.org/10.3758/s13415-019-00723-1>
- Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magnetic Resonance in Medicine*, 33(5), 636–647. <https://doi.org/10.1002/mrm.1910330508>
- Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, 12(8), 1062–1068. <https://doi.org/10.1038/nn.2342>
- Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience*, 35(2), 485–494. <https://doi.org/10.1523/jneurosci.2036-14.2015>
- Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, 104(41), 16311–16316. <https://doi.org/10.1073/pnas.0706111104>
- Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, 306(5703), 1940–1943. <https://doi.org/10.1126/science.1102941>
- Franklin, N. T., & Frank, M. J. (2018). Compositional clustering in task structure learning. *PLoS Computational Biology*, 14(4), e1006116. <https://doi.org/10.1371/journal.pcbi.1006116>
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4), 189–210. <https://doi.org/10.1002/hbm.460020402>
- Gardner, M. P., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B: Biological Sciences*, 285(1891), 20181645. <https://doi.org/10.1098/rspb.2018.1645>
- Garrett, H. E. (1922). A study of the relation of accuracy and speed. *Archives of Psychology*, 56.
- Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, 37(7), 1297–1310. <https://doi.org/10.1016/j.neubiorev.2013.03.023>
- Gerraty, R. T., Davidow, J. Y., Foerde, K., Galvan, A., Bassett, D. S., & Shohamy, D. (2018). Dynamic flexibility in striatal-cortical circuits supports reinforcement learning. *Journal of Neuroscience*, 38(10), 2442–2453. <https://doi.org/10.1523/jneurosci.2084-17.2018>
- Gerraty, R. T., Davidow, J. Y., Wimmer, G. E., Kahn, I., & Shohamy, D. (2014). Transfer of learning relates to intrinsic connectivity between hippocampus, ventromedial prefrontal cortex, and large-scale networks. *Journal of Neuroscience*, 34(34), 11297–11303. <https://doi.org/10.1523/jneurosci.0185-14.2014>
- Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, 71, 1–6. <https://doi.org/10.1016/j.jmp.2016.01.006>

- Gershman, S. J. (2017). Context-dependent learning and causal structure. *Psychonomic Bulletin & Review*, 24(2), 557–565. <https://doi.org/10.3758/s13423-016-1110-x>
- Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, 173, 34–42. <https://doi.org/10.1016/j.cognition.2017.12.014>
- Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective reevaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, 143(1), 182–194. <https://doi.org/10.1037/a0030844>
- Gershman, S. J., & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in Cognitive Science*, 7(3), 391–415. <https://doi.org/10.1111/tops.12138>
- Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1), 9–42. https://doi.org/10.1142/9789812797605_0002
- Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, 66(1), 15–36. <https://doi.org/10.1006/anbe.2003.2174>
- Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4), 585–595. <https://doi.org/10.1016/j.neuron.2010.04.016>
- Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*, 19(2), 483–495. <https://doi.org/10.1093/cercor/bhn098>
- Glover, G. H., Li, T. Q., & Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic Resonance in Medicine*, 44(1), 162–167. [https://doi.org/10.1002/1522-2594\(200007\)44:1%3c162::aid-mrm23%3e3.0.co;2-e](https://doi.org/10.1002/1522-2594(200007)44:1%3c162::aid-mrm23%3e3.0.co;2-e)
- Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30, 535–574. <https://doi.org/10.1146/annurev.neuro.29.051605.113038>
- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, 5, 13. <https://doi.org/10.3389/fninf.2011.00013>
- Grafton, S. T., Mazziotta, J. C., Woods, R. P., & Phelps, M. E. (1992). Human functional anatomy of visually guided finger movements. *Brain*, 115(2), 565–587. <https://doi.org/10.1093/brain/115.2.565>
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.
- Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1), 63–72. <https://doi.org/10.1016/j.neuroimage.2009.06.060>
- Grinband, J., Savitskaya, J., Wager, T. D., Teichert, T., Ferrera, V. P., & Hirsch, J. (2011). The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. *NeuroImage*, 57(2), 303–311. <https://doi.org/10.1016/j.neuroimage.2010.12.027>
- Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., & Haase, A. (2002). Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47(6), 1202–1210. <https://doi.org/10.1002/mrm.10171>
- Hadjikhani, N., Liu, A. K., Dale, A. M., Cavanagh, P., & Tootell, R. B. H. (1998). Retinotopy and color sensitivity in human visual cortical area V8. *Nature Neuroscience*, 1(3), 235–241. <https://doi.org/10.1038/681>
- Hampton, A. N., Adolphs, R., Tyszka, J. M., & O'Doherty, J. P. (2007). Contributions of the amygdala to reward expectancy and choice signals in human prefrontal cortex. *Neuron*, 55(4), 545–555. <https://doi.org/10.1016/j.neuron.2007.07.022>
- Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, 26(32), 8360–8367. <https://doi.org/10.1523/jneurosci.1010-06.2006>
- Hamrick, J. B., Ballard, A. J., Pascanu, R., Vinyals, O., Heess, N., & Battaglia, P. W. (2017). Metacognition for adaptive imagination-based optimization. *arXiv*, 1705.02670. <https://doi.org/10.48550/arxiv.1705.02670>
- Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience*, 28(22), 5623–5630. <https://doi.org/10.1523/jneurosci.1309-08.2008>
- Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P., & Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences*, 108(44), 18120–18125. <https://doi.org/10.1073/pnas.1109322108>
- Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, 56(1), 51–65. <https://doi.org/10.1037/h0062474>
- Harvey, A. K., Pattinson, K. T., Brooks, J. C., Mayhew, S. D., Jenkinson, M., & Wise, R. G. (2008). Brainstem functional magnetic resonance imaging: Disentangling signal from physiological noise. *Journal of Magnetic Resonance Imaging*, 28(6), 1337–1344. <https://doi.org/10.1002/jmri.21623>
- Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *NeuroImage*, 84, 159–168. <https://doi.org/10.1016/j.neuroimage.2013.08.028>
- Hauser, T. U., Iannaccone, R., Walitza, S., Brandeis, D., & Brem, S. (2015). Cognitive flexibility in adolescence: Neural and behavioral mechanisms of reward prediction error processing in adaptive decision making during development. *NeuroImage*, 104, 347–354. <https://doi.org/10.1016/j.neuroimage.2014.09.018>
- Horga, G., Maia, T. V., Marsh, R., Hao, X., Xu, D., Duan, Y., Tau, G. Z., Graniello, B., Wang, Z., Kangarlu, A., Martinez, D., Packard, M. G., & Peterson, B. S. (2015). Changes in corticostriatal connectivity during reinforcement learning in humans. *Human Brain Mapping*, 36(2), 793–803. <https://doi.org/10.1002/hbm.22665>
- Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297–307. <https://doi.org/10.1093/biomet/76.2.297>
- Ito, M., & Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience*, 29(31), 9861–9874. <https://doi.org/10.1523/jneurosci.6157-08.2009>
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79–87. <https://doi.org/10.1162/neco.1991.3.1.79>
- Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2), 825–841. <https://doi.org/10.1006/nimg.2002.1132>
- Jocham, G., Brodersen, K. H., Constantinescu, A. O., Kahn, M. C., Ianni, A. M., Walton, M. E., Rushworth, M. F. S., & Behrens, T. E. J. (2016). Reward-guided learning with and without causal attribution. *Neuron*, 90(1), 177–190. <https://doi.org/10.1016/j.neuron.2016.02.018>
- Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *Journal of Neuroscience*, 31(5), 1606–1613. <https://doi.org/10.1523/jneurosci.3904-10.2011>
- Johnson, D. M. (1939). Confidence and speed in the two-category judgment. *Archives of psychology*, 241.
- Jordan, M. I. (Ed.). (1998). *Learning in graphical models*. Springer. <https://doi.org/10.1007/978-94-011-5014-9>

- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. <https://doi.org/10.2307/1914185>
- Kahnt, T., Park, S. Q., Burke, C. J., & Tobler, P. N. (2012). How glitter relates to gold: Similarity-dependent reward prediction errors in the human striatum. *Journal of Neuroscience*, 32(46), 16521–16529. <https://doi.org/10.1523/jneurosci.2383-12.2012>
- Kahnt, T., & Tobler, P. N. (2016). Dopamine regulates stimulus generalization in the human hippocampus. *eLife*, 5, e12678. <https://doi.org/10.7554/elife.12678>
- Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, 3(8), 759–763. <https://doi.org/10.1038/77664>
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11), 4302–4311. <https://doi.org/10.1523/jneurosci.17-11-04302.1997>
- Karagoz, A., Reagh, Z., & Kool, W. (2022). The construction and use of cognitive maps in model-based control. *PsyArXiv*, ngqwa. <https://doi.org/10.31234/osf.io/ngqwa>
- Kasper, L., Bollmann, S., Diaconescu, A. O., Hutton, C., Heinze, J., Iglesias, S., Hauser, T. U., Sebold, M., Manjaly, Z., Pruessmann, K. P., & Stephan, K. E. (2017). The PhysIO toolbox for modeling physiological noise in fMRI data. *Journal of Neuroscience Methods*, 276, 56–72. <https://doi.org/10.1016/j.jneumeth.2016.10.019>
- Katahira, K. (2015). The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *Journal of Mathematical Psychology*, 66, 59–69. <https://doi.org/10.1016/j.jmp.2015.03.006>
- Katahira, K. (2018). The statistical structures of reinforcement learning with asymmetric value updates. *Journal of Mathematical Psychology*, 87, 31–45. <https://doi.org/10.1016/j.jmp.2018.09.002>
- Kato, A., & Morita, K. (2016). Forgetting in reinforcement learning links sustained dopamine signals to motivation. *PLoS Computational Biology*, 12(10), e1005145. <https://doi.org/10.1371/journal.pcbi.1005145>
- Kim, H., Shimojo, S., & O'Doherty, J. P. (2006). Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biology*, 4(8), e233. <https://doi.org/10.1371/journal.pbio.0040233>
- Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E. C., & Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLoS Computational Biology*, 13(2), e1005350. <https://doi.org/10.1371/journal.pcbi.1005350>
- Klopf, A. H. (1972). *Brain function and adaptive systems—A heterostatic theory* (Technical Report No. AFCRL-72-0164). Air Force Cambridge Research Laboratories.
- Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS Computational Biology*, 12(8), e1005090. <https://doi.org/10.1371/journal.pcbi.1005090>
- Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, 28(9), 1321–1333. <https://doi.org/10.1177/0956797617708288>
- Kool, W., Gershman, S. J., & Cushman, F. A. (2018). Planning complexity registers as a cost in metacontrol. *Journal of Cognitive Neuroscience*, 30(10), 1391–1404. https://doi.org/10.1162/jocn_a_01263
- Kurth-Nelson, Z., Economides, M., Dolan, R. J., & Dayan, P. (2016). Fast sequences of non-spatial state representations in humans. *Neuron*, 91(1), 194–204. <https://doi.org/10.1016/j.neuron.2016.05.028>
- Laming, D. R. J. (1968). *Information theory of choice-reaction times*. Academic Press.
- Langdon, A. J., Sharpe, M. J., Schoenbaum, G., & Niv, Y. (2018). Model-based predictions for dopamine. *Current Opinion in Neurobiology*, 49, 1–7. <https://doi.org/10.1016/j.conb.2017.10.006>
- Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, 84(3), 555–579. <https://doi.org/10.1901/jeab.2005.110-04>
- Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3), 687–699. <https://doi.org/10.1016/j.neuron.2013.11.028>
- Lehnert, L., Littman, M. L., & Frank, M. J. (2020). Reward-predictive representations generalize across tasks in reinforcement learning. *PLoS Computational Biology*, 16(10), e1008317. <https://doi.org/10.1371/journal.pcbi.1008317>
- Lesage, E., & Verguts, T. (2021). Contextual overtraining accelerates habit formation in new stimuli. *PsyArXiv*, 7m6bh. <https://doi.org/10.31234/osf.io/7m6bh>
- Lesaint, F., Sigaud, O., Flagel, S. B., Robinson, T. E., & Khamassi, M. (2014). Modelling individual differences in the form of Pavlovian conditioned approach responses: A dual learning systems approach with factored representations. *PLoS Computational Biology*, 10(2), e1003466. <https://doi.org/10.1371/journal.pcbi.1003466>
- Lewin, K. (1935). *A dynamic theory of personality*. McGraw-Hill.
- Lewin, K. (1936). *Principles of topological psychology*. McGraw-Hill. <https://doi.org/10.1037/10019-000>
- Li, J., & Daw, N. D. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. *Journal of Neuroscience*, 31(14), 5504–5511. <https://doi.org/10.1523/jneurosci.6316-10.2011>
- Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, 14(10), 1250–1252. <https://doi.org/10.1038/nn.2904>
- Li, L., Walsh, T. J., & Littman, M. L. (2006). Towards a unified theory of state abstraction for MDPs. *International Symposium on Artificial Intelligence and Mathematics*, 9.
- Lieberman, M. D., & Cunningham, W. A. (2009). Type I and type II error concerns in fMRI research: Re-balancing the scale. *Social Cognitive and Affective Neuroscience*, 4(4), 423–428. <https://doi.org/10.1093/scan/nsp052>
- Lim, S. L., O'Doherty, J. P., & Rangel, A. (2013). Stimulus value signals in ventromedial PFC reflect the integration of attribute value signals computed in fusiform gyrus and posterior superior temporal gyrus. *Journal of Neuroscience*, 33(20), 8729–8741. <https://doi.org/10.1523/jneurosci.4809-12.2013>
- Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human replay spontaneously reorganizes experience. *Cell*, 178(3), 640–652. <https://doi.org/10.1016/j.cell.2019.06.012>
- Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D., & Dolan, R. J. (2021). Experience replay is associated with efficient nonlocal learning. *Science*, 372(6544), eabf1357. <https://doi.org/10.1126/science.abf1357>
- Lohrenz, T., McCabe, K., Camerer, C. F., & Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences*, 104(22), 9493–9498. <https://doi.org/10.1073/pnas.0608842104>
- Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley. <https://doi.org/10.1037/14396-000>
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195070019.001.0001>
- Luzardo, A., Alonso, E., & Mondragón, E. (2017). A Rescorla-Wagner drift-diffusion model of conditioning and timing. *PLoS Computational Biology*, 13(11), e1005796. <https://doi.org/10.1371/journal.pcbi.1005796>
- Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4), 1122–1135. <https://doi.org/10.3758/s13428-014-0532-5>
- Magrabi, A., Ludwig, V. U., Stoppel, C. M., Paschke, L. M., Wisniewski, D., Heekeren, H. R., & Walter, H. (2021). Dynamic computation of value signals via a common neural network in multi-attribute decision-

- making. *Social Cognitive and Affective Neuroscience*, nsab125. <https://doi.org/10.1093/scan/nsab125>
- Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9), 1498–1507. <https://doi.org/10.1162/jocn.2007.19.9.1498>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman. <https://doi.org/10.7551/mitpress/9780262514620.001.0001>
- Masoudnia, S., & Ebrahimpour, R. (2014). Mixture of experts: A literature survey. *Artificial Intelligence Review*, 42(2), 275–293. <https://doi.org/10.1007/s10462-012-9338-y>
- Matsumoto, M., Matsumoto, K., Abe, H., & Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. *Nature Neuroscience*, 10(5), 647–656. <https://doi.org/10.1038/nn1890>
- Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21(11), 1609–1617. <https://doi.org/10.1126/science.abf1357>
- Mattar, M. G., Thompson-Schill, S. L., & Bassett, D. S. (2018). The network architecture of value learning. *Network Neuroscience*, 2(2), 128–149. https://doi.org/10.1162/netn_a_00021
- McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, 38(2), 339–346. [https://doi.org/10.1016/s0896-6273\(03\)00154-5](https://doi.org/10.1016/s0896-6273(03)00154-5)
- McDougle, S. D., & Collins, A. G. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, 28, 20–39. <https://doi.org/10.3758/s13423-020-01774-z>
- Miletić, S., Boag, R. J., & Forstmann, B. U. (2020). Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia*, 136, 107261. <https://doi.org/10.1016/j.neuropsychologia.2019.107261>
- Miletić, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks. *eLife*, 10, e63055. <https://doi.org/10.7554/elife.63055>
- Millner, A. J., Gershman, S. J., Nock, M. K., & den Ouden, H. E. (2018). Pavlovian control of escape and avoidance. *Journal of Cognitive Neuroscience*, 30(10), 1379–1390. https://doi.org/10.1162/jocn_a_01224
- Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., & Ugurbil, K. (2010). Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine*, 63(5), 1144–1153. <https://doi.org/10.1002/mrm.22361>
- Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, 7, e32548. <https://doi.org/10.7554/eLife.32548>
- Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behavior*, 1(9), 680–692. <https://doi.org/10.1038/s41562-017-0180-8>
- Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16(5), 1936–1947. <https://doi.org/10.1523/jneurosci.16-05-01936.1996>
- Montague, P. R., King-Casas, B., & Cohen, J. D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, 29, 417–448. <https://doi.org/10.1146/annurev.neuro.29.051605.112903>
- Morita, K., & Kato, A. (2014). Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. *Frontiers in Neural Circuits*, 8, 36. <https://doi.org/10.3389/fncir.2014.00036>
- Morris, L. S., Kundu, P., Costi, S., Collins, A., Schneider, M., Verma, G., Balchandani, P., & Murrugh, J. W. (2019). Ultra-high field MRI reveals mood-related circuit disturbances in depression: A comparison between 3-tesla and 7-tesla. *Translational Psychiatry*, 9(1), 1–11. <https://doi.org/10.1038/s41398-019-0425-6>
- Moser, E. I., Kropff, E., & Moser, M. B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, 31, 69–89. <https://doi.org/10.1146/annurev.neuro.31.061307.090723>
- Mumford, J. A., Poline, J. B., & Poldrack, R. A. (2015). Orthogonalization of regressors in fMRI models. *PLoS One*, 10(4), e0126255. <https://doi.org/10.1371/journal.pone.0126255>
- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44(1), 190–204. <https://doi.org/10.1006/jmps.1999.1283>
- Nakahara, H. (2014). Multiplexing signals in reinforcement learning with internal models and dopamine. *Current Opinion in Neurobiology*, 25, 123–129. <https://doi.org/10.1016/j.conb.2014.01.001>
- Nakahara, H., & Hikosaka, O. (2012). Learning to represent reward structure: A key to adapting to complex environments. *Neuroscience Research*, 74(3–4), 177–183. <https://doi.org/10.1016/j.neures.2012.09.007>
- Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7(4), 308–313. <https://doi.org/10.1093/comjnl/7.4.308>
- Newsome, W. T., & Paré, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience*, 8(6), 2201–2211. <https://doi.org/10.1523/jneurosci.08-06-02201.1988>
- Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2), 551–562. <https://doi.org/10.1523/jneurosci.5498-10.2012>
- Norbury, A., Robbins, T. W., & Seymour, B. (2018). Value generalization in human avoidance learning. *eLife*, 7, e34779. <https://doi.org/10.7554/elife.34779>
- O'Doherty, J. P. (2012). Beyond simple reinforcement learning: The computational neurobiology of reward-learning and valuation. *European Journal of Neuroscience*, 35(7), 987–990. <https://doi.org/10.1111/j.1460-9568.2012.08074.x>
- O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual Review of Psychology*, 68, 73–100. <https://doi.org/10.1146/annurev-psych-010416-044216>
- O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2), 329–337. [https://doi.org/10.1016/s0896-6273\(03\)00169-7](https://doi.org/10.1016/s0896-6273(03)00169-7)
- O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669), 452–454. <https://doi.org/10.1126/science.1094285>
- O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, 1104(1), 35–53. <https://doi.org/10.1196/annals.1390.022>
- O'Doherty, J. P., Lee, S., Tadayonnejad, R., Cockburn, J., Iigaya, K., & Charpentier, C. J. (2021). Why and how the brain weights contributions from a mixture of experts. *Neuroscience & Biobehavioral Reviews*, 123, 14–23. <https://doi.org/10.1016/j.neubiorev.2020.10.022>
- O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford University Press.
- Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66. <https://doi.org/10.1109/tsmc.1979.4310076>
- Palminteri, S. (2021). Choice-confirmation bias and gradual perseveration in human reinforcement learning. *PsyArXiv*, dpqj6. <https://doi.org/10.31234/osf.io/dpqj6>

- Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, 6(1), 1–14. <https://doi.org/10.1038/ncomms9096>
- Palminteri, S., & Lebreton, M. (2021). Context-dependent outcome encoding in human reinforcement learning. *Current Opinion in Behavioral Sciences*, 41, 144–151. <https://doi.org/10.1016/j.cobeha.2021.06.006>
- Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, 13(8), e1005684. <https://doi.org/10.1371/journal.pcbi.1005684>
- Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6), 425–433. <https://doi.org/10.1016/j.tics.2017.03.011>
- Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). Map making: Constructing, combining, and inferring on abstract cognitive maps. *Neuron*, 107(6), 1226–1238. <https://doi.org/10.1016/j.neuron.2020.06.030>
- Pauli, W. M., Larsen, T., Collette, S., Tyszka, J. M., Seymour, B., & O'Doherty, J. P. (2015). Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning. *Journal of Neuroscience*, 35(42), 14220–14233. <https://doi.org/10.1523/jneurosci.2277-15.2015>
- Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press.
- Pedersen, M. L., & Frank, M. J. (2020). Simultaneous hierarchical Bayesian parameter estimation for reinforcement learning and drift diffusion models: A tutorial and links to neural data. *Computational Brain & Behavior*, 3, 458–471. <https://doi.org/10.1007/s42113-020-00084-w>
- Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, 24(4), 1234–1251. <https://doi.org/10.3758/s13423-016-1199-y>
- Philiastides, M. G., Biele, G., & Heekeren, H. R. (2010). A mechanistic account of value computation in the human brain. *Proceedings of the National Academy of Sciences*, 107(20), 9430–9435. <https://doi.org/10.1073/pnas.1001732107>
- Pischedda, D., Palminteri, S., & Coricelli, G. (2020). The effect of counterfactual information on outcome value coding in medial prefrontal and cingulate cortex: From an absolute to a relative neural code. *Journal of Neuroscience*, 40(16), 3268–3277. <https://doi.org/10.1523/jneurosci.1712-19.2020>
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, 59(3), 2142–2154. <https://doi.org/10.1016/j.neuroimage.2011.10.018>
- Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, 84, 320–341. <https://doi.org/10.1016/j.neuroimage.2013.08.048>
- Prévost, C., McNamee, D., Jessup, R. K., Bossaerts, P., & O'Doherty, J. P. (2013). Evidence for model-based computations in the human amygdala during Pavlovian conditioning. *PLoS Computational Biology*, 9(2), e1002918. <https://doi.org/10.1371/journal.pcbi.1002918>
- Pruim, R. H. R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage*, 112, 267–277. <https://doi.org/10.1016/j.neuroimage.2015.02.064>
- Rangel, A., & Clithero, J. A. (2012). Value normalization in decision making: Theory and evidence. *Current Opinion in Neurobiology*, 22(6), 970–981. <https://doi.org/10.1016/j.conb.2012.07.011>
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59–108. <https://doi.org/10.1037/0033-295x.85.2.59>
- Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and diffusion models. *Neural Computation*, 24(5), 1186–1229. https://doi.org/10.1162/neco_a_00270
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20(4), 260–281. <https://doi.org/10.1016/j.tics.2016.01.007>
- Reiter, A. M. F., Heinze, H., Schlagenhaut, F., & Deserno, L. (2017). Impaired flexible reward-based decision-making in binge eating disorder: Evidence from computational modeling and functional neuroimaging. *Neuropsychopharmacology*, 42(3), 628–637. <https://doi.org/10.1038/npp.2016.95>
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.
- Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems* (Technical report no. CUED/F-INFENG/TR 166). Department of Engineering, University of Cambridge.
- Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, 13(9), e1005768. <https://doi.org/10.1371/journal.pcbi.1005768>
- Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2021). Neural evidence for the successor representation in choice evaluation. *bioRxiv*, 458114. <https://doi.org/10.1101/2021.08.29.458114>
- Sadacca, B. F., Jones, J. L., & Schoenbaum, G. (2016). Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife*, 5, e13665. <https://doi.org/10.7554/elife.13665.001>
- Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, 27(47), 12860–12867. <https://doi.org/10.1523/jneurosci.2496-07.2007>
- Schuck, N. W., & Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science*, 364(6447), eaaw5181. <https://doi.org/10.1126/science.aaw5181>
- Schultz, W. (2013). Updating dopamine reward signals. *Current Opinion in Neurobiology*, 23(2), 229–238. <https://doi.org/10.1016/j.conb.2012.11.012>
- Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Reviews*, 95(3), 853–951. <https://doi.org/10.1152/physrev.00023.2014>
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599. <https://doi.org/10.1126/science.275.5306.1593>
- Schulz, E., Franklin, N. T., & Gershman, S. J. (2020). Finding structure in multi-armed bandits. *Cognitive Psychology*, 119, 101261. <https://doi.org/10.1016/j.cogpsych.2019.101261>
- Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, 55, 7–14. <https://doi.org/10.1016/j.conb.2018.11.003>
- Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2018). Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(6), 927–943. <https://doi.org/10.1037/xlm0000463>
- Sengupta, A., Speck, O., Yakupov, R., Kanowski, M., Tempelmann, C., Pollmann, S., & Hanke, M. (2018). The effect of acquisition resolution on orientation decoding from V1: Comparison of 3T and 7T. *bioRxiv*, 305417. <https://doi.org/10.1101/305417>
- Sewell, D. K., Jach, H. K., Boag, R. J., & Van Heer, C. A. (2019). Combining error-driven models of associative learning with evidence accumulation models of decision-making. *Psychonomic Bulletin & Review*, 26(3), 868–893. <https://doi.org/10.3758/s13423-019-01570-4>

- Sewell, D. K., & Stallman, A. (2020). Modeling the effect of speed emphasis in probabilistic category learning. *Computational Brain & Behavior*, 3(2), 129–152. <https://doi.org/10.1007/s42113-019-00067-6>
- Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N. S. P. N., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Computational Biology*, 15(2), e1006803. <https://doi.org/10.1371/journal.pcbi.1006803>
- Sharot, T. (2011). The optimism bias. *Current Biology*, 21(23), R941–R945. <https://doi.org/10.1016/j.cub.2011.10.030>
- Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, 14(11), 1475–1479. <https://doi.org/10.1038/nn.2949>
- Shenhav, A., Straccia, M. A., Cohen, J. D., & Botvinick, M. M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature Neuroscience*, 17(9), 1249–1254. <https://doi.org/10.1038/nn.3771>
- Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22(4), 325–345. <https://doi.org/10.1007/bf02288967>
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820), 1317–1323. <https://doi.org/10.1126/science.3629243>
- Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(Supplement 1), S208–S219. <https://doi.org/10.1016/j.neuroimage.2004.07.051>
- Smyser, C. D., Inder, T. E., Shimony, J. S., Hill, J. E., Degnan, A. J., Snyder, A. Z., & Neil, J. J. (2010). Longitudinal analysis of neural network development in preterm infants. *Cerebral Cortex*, 20(12), 2852–2862. <https://doi.org/10.1093/cercor/bhq035>
- Soellinger, M., Ryf, S., Boesiger, P., & Kozerke, S. (2007). Assessment of human brain motion using CSPAMM. *Journal of Magnetic Resonance Imaging*, 25(4), 709–714. <https://doi.org/10.1002/jmri.20882>
- Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, 7(2), 351–367. <https://doi.org/10.1111/tops.12145>
- Stojić, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2020). It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *Journal of Experimental Psychology: General*, 149, 1878–1907. <https://doi.org/10.1037/xge0000749>
- Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, 25(3), 251–260. <https://doi.org/10.1007/bf02289729>
- Sugawara, M., & Katahira, K. (2021). Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Scientific Reports*, 11, 3574. <https://doi.org/10.1038/s41598-020-80593-7>
- Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning* (Doctoral dissertation). University of Massachusetts, Amherst.
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1), 9–44. <https://doi.org/10.1007/bf00115009>
- Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In B. W. Porter & R. J. Mooney (Eds.), *Machine learning: Proceedings of the seventh international conference* (pp. 216–224). Morgan Kaufmann. <https://doi.org/10.1016/b978-1-55860-141-3.50030-4>
- Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4), 160–163. <https://doi.org/10.1145/122344.122377>
- Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88(2), 135–170. <https://doi.org/10.1037/0033-295X.88.2.135>
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4), 629–640. <https://doi.org/10.1017/s0140525x01000061>
- Theysohn, N., Qin, S., Maderwald, S., Poser, B. A., Theysohn, J. M., Ladd, M. E., Norris, D. G., Gizewski, E. R., Fernandez, G., & Tendolcar, I. (2013). Memory-related hippocampal activity can be measured robustly using fMRI at 7 tesla. *Journal of Neuroimaging*, 23(4), 445–451. <https://doi.org/10.1111/jon.12036>
- Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4), 285–294. <https://doi.org/10.2307/2332286>
- Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review: Series of Monograph Supplements*, 2, 4. <https://doi.org/10.5962/bhl.title.25848>
- Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Macmillan. <https://doi.org/10.5962/bhl.title.55072>
- Thorndike, E. L. (1932). *The fundamentals of learning*. Teachers College Bureau of Publications, Columbia University. <https://doi.org/10.1037/10976-000>
- Thorndike, E. L. (1933). A proof of the law of effect. *Science*, 77(1989), 173–175. <https://doi.org/10.1126/science.77.1989.173-a>
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189–208. <https://doi.org/10.1037/h0061626>
- Tomov, M. S., Dorfman, H. M., & Gershman, S. J. (2018). Neural computations underlying causal structure learning. *Journal of Neuroscience*, 38(32), 7143–7157. <https://doi.org/10.1523/jneurosci.3336-17.2018>
- Tootell, R. B. H., Reppas, J. B., Kwong, K. K., Malach, R., Born, R. T., Brady, T. J., Rosen, B. R., & Belliveau, J. W. (1995). Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. *Journal of Neuroscience*, 15(4), 3215–3230. <https://doi.org/10.1523/jneurosci.15-04-03215.1995>
- Torrì, S., Chen, G., Glen, D., Bandettini, P. A., Baker, C. I., Reynolds, R., Liu, J. Y., Leshin, J., Balderston, N., Grillon, C., & Ernst, M. (2018). Statistical power comparisons at 3T and 7T with a GO/NOGO task. *NeuroImage*, 175, 100–110. <https://doi.org/10.1016/j.neuroimage.2018.03.071>
- Toyama, A., Katahira, K., & Ohira, H. (2017). A simple computational algorithm of model-based choice preference. *Cognitive, Affective, & Behavioral Neuroscience*, 17(4), 764–783. <https://doi.org/10.3758/s13415-017-0511-2>
- Toyama, A., Katahira, K., & Ohira, H. (2019). Reinforcement learning with parsimonious computation and a forgetting process. *Frontiers in Human Neuroscience*, 13, 153. <https://doi.org/10.3389/fnhum.2019.00153>
- Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, 29(11), 2225–2232. <https://doi.org/10.1111/j.1460-9568.2009.06796.x>
- Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6), 1310–1320. <https://doi.org/10.1109/tmi.2010.2046908>
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–352. <https://doi.org/10.1037/0033-295x.84.4.327>
- Uğurbil, K. (2018). Imaging at ultrahigh magnetic fields: History, challenges, and solutions. *NeuroImage*, 168, 7–32. <https://doi.org/10.1016/j.neuroimage.2017.07.007>
- Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3), 550–592. <https://doi.org/10.1037/0033-295x.108.3.550>
- van Dam, L. C., & Ernst, M. O. (2015). Mapping shape to visuomotor mapping: Learning and generalisation of sensorimotor behaviour based on contextual information. *PLoS Computational Biology*, 11(3), e1004172. <https://doi.org/10.1371/journal.pcbi.1004172>

- Viejo, G., Khamassi, M., Brovelli, A., & Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in Behavioral Neuroscience*, 9, 225. <https://doi.org/10.3389/fnbeh.2015.00225>
- Vinckier, F., Gaillard, R., Palminteri, S., Rigoux, L., Salvador, A., Fornito, A., Adapa, R., Krebs, M. O., Pessiglione, M., & Fletcher, P. C. (2016). Confidence and psychosis: A neuro-computational account of contingency learning disruption by NMDA blockade. *Molecular Psychiatry*, 21(7), 946–955. <https://doi.org/10.1038/mp.2015.73>
- Wade, A. R., Brewer, A. A., Rieger, J. W., & Wandell, B. A. (2002). Functional measurements of human ventral occipital cortex: Retinotopy and colour. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 357(1424), 963–973. <https://doi.org/10.1098/rstb.2002.1108>
- Wang, X. J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5), 955–968. [https://doi.org/10.1016/s0896-6273\(02\)01092-9](https://doi.org/10.1016/s0896-6273(02)01092-9)
- Watanabe, K., & Hikosaka, O. (2005). Immediate changes in anticipatory activity of caudate neurons associated with reversal of position-reward contingency. *Journal of Neurophysiology*, 94(3), 1879–1887. <https://doi.org/10.1152/jn.00012.2005>
- Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Doctoral dissertation). University of Cambridge.
- Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3–4), 279–292. <https://doi.org/10.1007/bf00992698>
- Watson, J. D. G., Myers, R., Frackowiak, R. S. J., Hajnal, J. V., Woods, R. P., Mazziotta, J. C., Shipp, S., & Zeki, S. (1993). Area V5 of the human brain: Evidence from a combined study using positron emission tomography and magnetic resonance imaging. *Cerebral Cortex*, 3(2), 79–94. <https://doi.org/10.1093/cercor/3.2.79>
- Weissman, D. H., & Carp, J. (2013). The congruency effect in the posterior medial frontal cortex is more consistent with time on task than with response conflict. *PLoS One*, 8(4), e62405. <https://doi.org/10.1371/journal.pone.0062405>
- Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, 8, e49547. <https://doi.org/10.7554/elife.49547>
- Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General*, 143(6), 2074–2081. <https://doi.org/10.1037/a0038199>
- Wimmer, G. E., Daw, N. D., & Shohamy, D. (2012). Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience*, 35(7), 1092–1104. <https://doi.org/10.1111/j.1460-9568.2012.08017.x>
- Wimmer, G. E., Liu, Y., Vejar, N., Behrens, T. E. J., & Dolan, R. J. (2020). Episodic memory retrieval success is associated with rapid replay of episode content. *Nature Neuroscience*, 23(8), 1025–1033. <https://doi.org/10.1038/s41593-020-0649-z>
- Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science*, 338(6104), 270–273. <https://doi.org/10.1126/science.1223252>
- Witten, I. H. (1977). An adaptive optimal controller for discrete-time Markov environments. *Information and Control*, 34(4), 286–295. [https://doi.org/10.1016/s0019-9958\(77\)90354-0](https://doi.org/10.1016/s0019-9958(77)90354-0)
- Wong, K. F., & Wang, X. J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4), 1314–1328. <https://doi.org/10.1523/jneurosci.3733-05.2006>
- Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2018). Connecting conceptual and spatial search via a model of generalization. *bioRxiv*, 258665. <https://doi.org/10.1101/258665>
- Wu, C. M., Schulz, E., & Gershman, S. J. (2019). Generalization as diffusion: Human function learning on graphs. *bioRxiv*, 538934. <https://doi.org/10.1101/538934>
- Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, 2(12), 915–924. <https://doi.org/10.1038/s41562-018-0467-4>
- Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, 15(5), 786–791. <https://doi.org/10.1038/nn.3068>
- Wunderlich, K., Symmonds, M., Bossaerts, P., & Dolan, R. J. (2011). Hedging your bets by learning reward correlations in the human brain. *Neuron*, 71(6), 1141–1152. <https://doi.org/10.1016/j.neuron.2011.07.025>
- Yarkoni, T., Barch, D. M., Gray, J. R., Conturo, T. E., & Braver, T. S. (2009). BOLD correlates of trial-by-trial reaction time variability in gray and white matter: A multi-study fMRI analysis. *PLoS One*, 4(1), e4257. <https://doi.org/10.1371/journal.pone.0004257>
- Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8), 665–670. <https://doi.org/10.1038/nmeth.1635>
- Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8), 1177–1193. <https://doi.org/10.1109/tnnls.2012.2200299>
- Zaki, J., Kallman, S., Wimmer, G. E., Ochsner, K., & Shohamy, D. (2016). Social cognition as reinforcement learning: Feedback modulates emotion inference. *Journal of Cognitive Neuroscience*, 28(9), 1270–1282. https://doi.org/10.1162/jocn_a_00978
- Zaretskaya, N., Bause, J., Polimeni, J. R., Grassi, P. R., Scheffler, K., & Bartels, A. (2020). Eye-selective fMRI activity in human primary visual cortex: Comparison between 3T and 9.4T, and effects across cortical depth. *NeuroImage*, 220, 117078. <https://doi.org/10.1016/j.neuroimage.2020.117078>
- Zeki, S., Watson, J. D. G., Lueck, C. J., Friston, K. J., Kennard, C., & Frackowiak, R. S. J. (1991). A direct demonstration of functional specialization in human visual cortex. *Journal of Neuroscience*, 11(3), 641–649. <https://doi.org/10.1523/jneurosci.11-03-00641.1991>
- Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: Frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, 15(6), 695–707. <https://doi.org/10.1093/scan/nsaa089>
- Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1), 45–57. <https://doi.org/10.1109/42.906424>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Colas, J. T., Dundon, N. M., Gerraty, R. T., Saragosa-Harris, N. M., Szymula, K. P., Tanwisuth, K., Tyszka, J. M., van Geen, C., Ju, H., Toga, A. W., Gold, J. I., Bassett, D. S., Hartley, C. A., Shohamy, D., Grafton, S. T., & O'Doherty, J. P. (2022). Reinforcement learning with associative or discriminative generalization across states and actions: fMRI at 3 T and 7 T. *Human Brain Mapping*, 43(15), 4750–4790. <https://doi.org/10.1002/hbm.25988>