Evidence for hierarchically-structured reinforcement learning in humans

Maria K Eckstein
Department of Psychology
UC Berkeley
Berkeley, CA 94720
maria.eckstein@berkeley.edu

Anne GE Collins
Department of Psychology
UC Berkeley
Berkeley, CA 94720
annecollins@berkeley.edu

Abstract

Flexibly adapting behavior to different contexts is a critical component of human intelligence. It requires knowledge to be structured as coherent, context-dependent action rules, or task-sets (TS). Nevertheless, inferring optimal TS is computationally complex. This paper tests the key predictions of a neurally-inspired model that employs hierarchically-structured reinforcement learning (RL) to approximate optimal inference. The model proposes that RL acts at two levels of abstraction: a high-level RL process learns context-TS values, which guide TS selection based on context; a low-level process learns stimulus-actions values within TS, which guide action selection in response to stimuli. In our novel task paradigm, we found evidence that participants indeed learned values at both levels: not only stimulus-action values, but also context-TS values affected learning and TS reactivation, and TS values alone determined TS generalization. This supports the claim of two RL processes, and their importance in structuring our interactions with the world.

Keywords: Reinforcement learning; structure learning; hierarchical representation; task sets

Introduction

Humans structure their knowledge of the world in a way that allows them to adapt to complex, ever-changing environments. Specifically, humans create different behavioral strategies (or "task sets", TS) for different contexts. For example, the TS of using the Mac operating system contains the set of behavioral rules that can be applied to Mac computers. More generally, contexts elicit specific TS, which in turn trigger the responses to environmental stimuli, a crucial function of cognitive control (Miller & Cohen, 2001). The Mac TS, for example, might be elicited by context cues such as Apple computers.

Previous research has employed models incorporating inference of latent structure to explain human learning and generalization in complex environments. For example, (Collins & Koechlin, 2012; Donoso, Collins, & Koechlin, 2014) used approximate Bayesian inference to capture how humans learn and select TS in different contexts. Human strategy selection approximates Bayes-optimal solutions for selecting context-appropriate TS, creating TS in new contexts, and assessing the reliability of current TS.

Nevertheless, the complexity of conducting Bayesian inference on latent variables such as TS makes it unlikely that this model provides a mechanistic description of human cognitive processes. Collins & Frank (2012) suggested that a hierarchically-structured reinforcement learning (RL) algorithm with a biologically plausible architecture underlies these processes. Crucially, they showed that such an algorithm can approximate Bayes-optimal inference using much simpler computations (Collins & Frank, 2013).

The hierarchical RL model makes several predictions about human behavior that go beyond the Bayesian model. The goal of the current study is to test some of these predictions. Most notably, the RL model predicts that humans select TS based on "TS values" (see below for details), and that these TS values influence behavior in several ways. Crucially, previous models based on Bayesian inference do not track TS values and thus predict no effects of TS values.

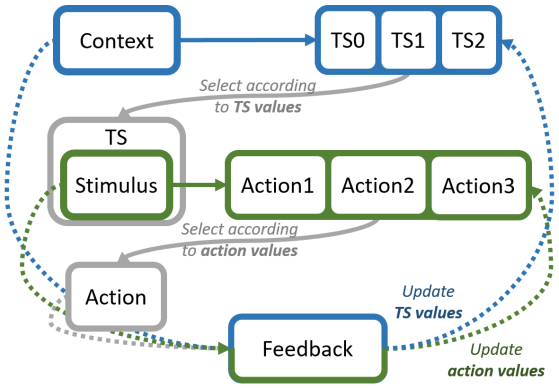

Figure 1: Schematic of the hierarchical RL model. The high-level loop (blue) selects TS based on TS values; the low-level loop (green) selects actions based on action values. TS values are based on context cues and action values are based on the selected TS and the current stimulus. TS and action values are learned over time from a continuous feedback signal that specifies the amount of reward obtained.

The hierarchical RL model relies on RL theory (Sutton & Barto, 2017). One basic principle of RL is the computation of stimulus-action values, which estimate how much cumulative future reward should be expected if an action is selected in response to a specific stimulus. A simple but reliable RL algorithm learns values by updating its estimates in proportion to a reward prediction error signal. Previous research has shown that RL algorithms provide good models for human and animal learning (Daw, Gershman, Seymour, Dayan, & Dolan, 2011) and capture important aspects of reward-based learning in the brain (Schultz, Dayan, & Montague, 1997), likely implemented in cortico-striatal loops (Alexander, DeLong, & Strick, 1986). However, in their simplest form, RL algorithms learn values independently for all stimulus-action pairs, and thus cannot account for generalization of learned

TS to new contexts. The hierarchical RL model tested here aims to integrate the strengths of both approaches in order to explain human reasoning in complex environments.

In the model, a low-level RL loop acquires associations between stimuli and actions by learning stimulus-action values through reinforcement (Fig. 1, green loop). Crucially, action values that are acquired in the same context are grouped into TS (coherent sets of stimulus-action mappings). Another, high-level loop learns associations between contexts and these TS by learning context-TS values (Fig. 1, blue loop). TS values guide the activation of TS in response to contexts. When a new context is encountered, the agent can either retrieve an old TS or create a new one, potentially informed by existing TS (see Collins & Koechlin, 2012, for an example of a similar model).

Taken together, the high-level loop in the hierarchical RL model influences the workings of the low-level loop by selecting the current TS, which determines action values. The low-level loop uses these action values to select actions. Crucially, both loops employ the same RL algorithms to compute values, and both are assumed to be implemented in similar neural substrate, cortico-striatal loops that only differ in their position on the anterior-posterior axis (Alexander et al., 1986). This model is supported both by knowledge about the brain's structural and functional organization (Alexander et al., 1986; Badre & Frank, 2012), and by computational models of human behavior (Frank & Badre, 2012). The model's key prediction is that humans perform RL at different levels in parallel, learning values at different levels of abstraction from a single feedback signal.

Here, we test qualitative predictions of this model. Future work will include quantitative model simulations, fitting and comparison; the current work focuses on behavioral analysis. First, we verify that participants create TS and flexibly reactivate them if needed. We then test whether participants acquire TS values and if these values affect behavior. Specifically, we predict that 1) TS values affect learning, such that higher-valued TS are acquired faster; 2) TS values influence context preferences, such that participants select higher-valued contexts when asked to choose; 3) TS are selected based on TS values, such that participants preferentially activate TS of higher values in new or unknown contexts; 4) TS values influence generalization, such that newly created TS are more similar to higher-valued TS. Our results support these predictions.

Methods

Current study

In order to test these predictions, we designed a reward-based associative learning task, in which participants encountered different contexts and learned the TS for each one. Contexts specified unique mappings between stimuli, responses, and outcomes, such that stimuli that were associated with high rewards in one context might be associated with small rewards in others (Fig. 2). After initial learning of the TS, participants underwent multiple testing phases, which aimed to test each of our predictions.

Task details

In our novel task, participants encountered four aliens and were asked to "help each one grow as much as possible". In each trial, participants saw one of four aliens along with three items. Participants selected one item by button press and received feedback as to how much the alien grew in response, indicated by the length of a measuring tape (reward). In each context, only one item led to a high reward for a given alien (correct action), whereas the other two items had similar small effects (incorrect actions). Therefore, each TS was specified by the correct mapping between each of the four aliens (stimulus) with one item (response).

Participants learned a different TS for each of three contexts (hot, cold, and rainy "seasons"; Fig. 2A), such that there was a one-to-one mapping between contexts and TS. Participants got 52 trials for each context (13 per alien) before encountering a new context, for a total of three repetitions per context during initial learning. The reward value associated with each correct context-alien-item mapping was normally distributed around a fixed mean, with standard deviation 0.5. The mean reward values were predetermined such that TS differed in average reward ("TS value"), while aliens and items did not (Fig. 2B). This manipulation allowed us to test participants' sensitivity to TS values, while ruling out confounds based on stimulus and action values.

The different phases of the alien task are described in table 1. To minimize confounds, the mappings between contexts and TS were randomized between participants, as were the images of aliens and items.

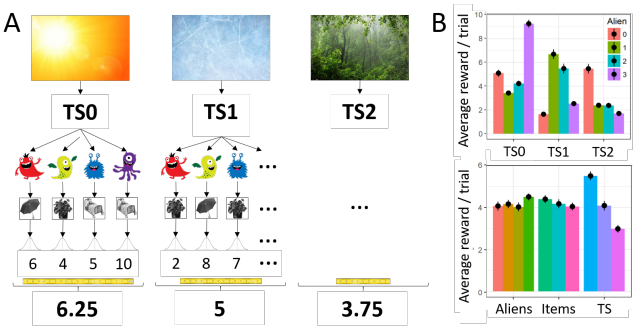

Figure 2: TS mappings and values at all levels. A) Three contexts (top row) were associated with three TS, as explained in the main text. B) Reward sizes differed between stimulus-response mappings (top), leading to differences in TS values, but not alien (stimulus) or item (action) values (bottom).

Participants

We tested 51 participants (26 women). One participant was excluded because the performance criterion of 50% was not reached in the practice round. The mean age was 22.1 years

Table 1: Description and purpose of the task phases.

Phase	Description	Purpose
Initial Learning (Phase 1)	Participants see one of four aliens at a time, select one item, and receive feedback; trials of the same context are presented in a block; context order is pseudo-randomized	Participants acquire different TS in each of three contexts, through trial and error
Refresher (Interleaved)	Similar to initial learning, but fewer trials; interleaved between testing phases	Restore TS, alleviate carry-over effects
Hidden Context (Phase 2)	Similar to initial learning, but current context is invisible; context changes are signaled	Test whether participants have acquired TS
Comparison (Phase 3)	Participants see two stimuli at a time and indicate their preferred one; two contexts, items, aliens, or context-alien combinations are presented at a time	Test whether participants learn TS values (high-level) and stimulus values (low-level)
Generalization (Phase 4)	Similar to initial learning, but in a novel context, without feedback	Test whether TS values affect generalization

(sd: 1.5 years). Participants were recruited from the UC Berkeley research participation pool and gave informed consent prior to participation.

TS are created and reactivated

Reactivation of TS in old contexts

We used the hidden-context phase to test whether participants learned coherent, context-specific TS, rather than non-hierarchical context-stimulus-response associations. In the hidden-context phase, participants were presented with a background of thick clouds instead of seeing the current context. The task still signaled when the context changed. After each context change, participants at first needed to guess which actions were correct because the new context was unknown. After a few trials, the received feedback allowed them to infer the context and to apply the correct TS (Collins & Koechlin, 2012).

A TS is a coherent, interdependent assembly of stimulus-response mappings that apply in a specific context. Because of the interdependence between mappings, certainty about some mappings should facilitate recall of the remaining mappings: for example, if participants successfully selected an umbrella for the red alien, they should infer that the context was "hot" and select the bed for the purple alien, before having observed this association (Fig. 2).

In order to test whether participants had formed TS, we first focused on trials in which participants saw an alien for the first time after a context change. In this situation, participants had not yet received any information about the correct item for this alien, i.e., they had no direct evidence about the stimulus-response mapping. We compared two different conditions within these trials, (1) when participants had not yet selected the correct item for any other alien, and (2) when participants had at least once selected the correct item for another alien. We expected that knowledge about some stimulus-response (alien-item) mappings within a TS would facilitate the recall of the remaining mappings, such that participants would select the correct item more often in condition (2) than (1). Note that this prediction is specific to models with latent structure, such as the Bayesian model proposed by Collins & Koechlin (2012) and our hierarchical RL model.

This was indeed the case. In condition (1), participants selected the correct item in 36.9% of trials, compared to 45.8% in condition (2) (chance: 33.3%; Fig. 3B). The difference was statistically significant (t(49)=2.5, p=0.014), suggesting that participants retrieved stimulus-response mappings for unseen aliens based on knowledge about already-seen alien-item mappings within the same TS. These results were confirmed in a regression model encompassing all trials of the hidden-context phase, rather than just the subset used above. In this model, each trial's accuracy was predicted from four factors, including (1) participants' performance in the previous trial of the same stimulus-response mapping ("ACC same"), and (2) participants' performance in the previous trials of the other three stimulus-response mappings combined ("ACC other"). As expected, both factors significantly affected performance (table 2).

Similar patterns were evident in the initial-learning phase and the two refresher periods. In these phases, the background pictures provided perfect cues for the current TS, as opposed to the hidden-context phase. The fact that certainty about other mappings (ACC other) still affected performance suggests that participants used partial knowledge about TS as a cue for the remaining mappings even when a perfect cue for the TS was given.

Taken together, the interdependence between stimulus-response mappings within contexts provides evidence that participants acquired coherent, stable, consistent TS. This replicates prior results (Collins & Frank, 2013) and is a precondition to test our novel predictions.

Transfer of TS to new contexts

Evidence for the reactivation of existing TS also comes from the generalization phase of our task. In this phase, participants were presented with the same four aliens, but in a novel context. Like before, participants were tasked with selecting the correct item for each alien, but no feedback was given, such that participants were continuously forced to guess.

We found that participants did not guess randomly, but instead reactivated prior TS. Items that were correct in a previously-learned TS were selected more often (90.5% of valid trials) than expected from random behavior (chance was 83.3%=10/12 because 10 out the 12 possible stimulus-response mappings were valid in at least one TS), t(49)=4.35, $p < 0.001$. This shows that when encountering novel contexts, participants reactivated old TS, rather than trying out novel stimulus-response mappings, in accordance with prior findings (Collins & Frank, 2013).

Table 2: Logistic mixed-effects regression predicting trial-wise accuracy from accuracy on the same (ACC same) and other mappings (ACC other).

Task phase	Predictor	β	p
Initial learning	**ACC same**	1.09	< 0.001
	ACC other	0.41	< 0.001
	interaction	0.31	0.028
	Repetition	0.18	<0.001
Refresher 1	**ACC same**	1.74	< 0.001
	ACC other	1.39	< 0.001
	interaction	-0.65	0.020
	Repetition	0.02	0.74
Refresher 2	**ACC same**	1.91	< 0.001
	ACC other	1.79	< 0.001
	interaction	-1.11	0.005
	Repetition	0.03	0.64
Hidden context	**ACC same**	1.18	< 0.001
	ACC other	0.72	< 0.001
	interaction	0.45	0.027
	Repetition	0.14	0.013

Sensitivity to TS values

So far, we have established that participants created TS and flexibly reactivated them when the context was hidden or novel. This replicates prior findings and is also predicted by non-RL models of latent learning. We next assessed whether and in which ways TS values affected behavior, a prediction that is novel and specific to RL-based models.

TS values affect learning

To this aim, we analyzed the initial-learning phase of the task, in which participants first learned to associate each alien with the correct item, in each context. To test the effects of both stimulus-action values (low-level RL loop) and context-TS values (high-level loop), we used a regression model predicting accuracy in each trial from four factors: (1) the value of the current stimulus (low-level), (2) the value of the current TS (high-level), (3) the trial index, in order to account for learning within a context block, and (4) the repetition index, which accounts for learning across context blocks (Fig. 3A). The model revealed a significant effect of stimulus values (table 3), as predicted by classic non-hierarchical RL models. This shows that participants performed better when correct responses were rewarded more. But the model also revealed an effect of TS values, after controlling for stimulus values. This additional influence of TS values goes beyond predictions of classic non-hierarchical RL models, but is predicted by our model.

TS values affect context selection

We next assessed whether TS values influenced context selection, such that participants would prefer contexts that had been associated with higher-valued TS to those associated with lower-valued TS. We tested this prediction in the comparison phase of the task. Here, participants were presented with the images of two different contexts and were asked to select their preferred one.

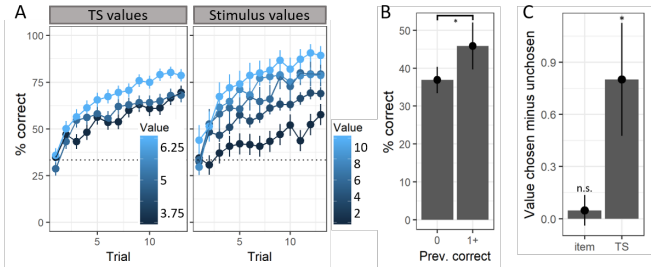

Figure 3: A) Influence of TS and stimulus values on performance in the initial-learning phase. B) Effect of performance in other mappings on performance in the current mapping, within a TS. Left: No prior correct responses in other mappings; right: at least one correct response. C) Effect of TS values on context selection. Value difference between chosen and unchosen items (left) and TS (right).

Table 3: Logistic mixed-effects regression predicting trial-wise accuracy from stimulus values, TS values, and trial index.

Task phase	Predictor	β	p
Initial learning	**Stimulus value**	0.19	< 0.001
	TS value	0.037	0.24
	Trial index	0.14	< 0.001
	Repetition	0.30	< 0.001
Refresher 1	**Stimulus value**	0.12	< 0.001
	TS value	0.18	< 0.001
	Trial index	0.31	< 0.001
	Repetition	0.16	0.053
Refresher 2	**Stimulus value**	0.12	< 0.001
	TS value	0.18	0.0026
	Trial index	0.27	< 0.001
	Repetition	0.22	0.019
Hidden context	**Stimulus value**	0.20	< 0.001
	TS value	0.09	0.048
	Trial index	0.27	< 0.001
	Repetition	0.22	0.0038

We calculated individual TS values for each participant, based on individual learning history. Participants indeed chose contexts more often that had been associated with higher-valued TS (68.9% of trials), t(49)=5.07, $p < 0.001$, resulting in a significant difference between TS values of chosen (4.56) and unchosen contexts (3.75), t(49)=5.00, $p < 0.001$ (Fig. 3C). This shows that participants indeed selected contexts based on the values of associated TS.

A similar pattern arose for action (item) values. We again calculated individual values. (Although items did not differ in their objective values, slight differences arose because of participants' individual decision histories.) Participants chose higher-valued items more often (62.5 % of trials), t(49)=3.47, $p = 0.001$, although the difference between chosen (4.14) and unchosen items (4.09) was not significant, t(49)=1.09, $p =$ 28, presumably because of the lack of spread in item values (Fig. 3C). The first result still implies that participants had learned action values in addition to TS values.

We also aimed to confirm that participants had acquired traditional stimulus values. To test this, we asked participants to select between two aliens in the same context. We expected that participants would prefer the aliens that were associated with larger rewards, as has been shown many times before (Frank, Seeberger, & O'Reilly, 2004). Unfortunately, due to a technical error, we were unable to confirm this here. Overall, our results imply that participants learned different sets of values, as predicted by our model. High-level TS values influenced TS learning and guided TS selection. Low-level stimulus values affected stimulus-response learning and are expected to affect stimulus selection.

TS values affect generalization

Having shown that participants learned values for TS and that TS values affected learning and context selection, we next tested whether TS values also affected TS reactivation and generalization. We tested this in three ways. First, we asked whether higher-valued TS were reactivated more readily than lower-valued TS in the hidden-context phase. Second, we assessed whether TS values predicted error types in the initial-learning phase and the refreshers. Third, we tested whether higher-valued TS had a larger influence on the creation of new TS than lower-valued ones (generalization phase).

With respect to our first question, we have shown above that TS values influenced accuracy in the hidden-context phase (table 3). One potential reason for this is that participants more readily reactivated higher-valued than lower-valued TS, leading to higher accuracy in higher-valued TS as a whole. This is in accordance with our model, which predicts that TS are selected based on TS values.

Our second assessment concerned errors, specifically intrusions from other TS. We defined intrusion as the selection of an action that is correct in a context other than the current one. In the initial-learning phase, if participants made errors by selecting incorrect items uniformly at random, 75% of all errors would be intrusions (due to the specific way TS were defined). Participants instead produced 78.4% intrusion errors, a small but significant increase, t(49)=6.69, $p < 0.001$. Within these intrusions, TS values significantly affected item selection, as shown in a logistic mixed-effects regression model (table 4). The effect was not driven by stimulus values. This confirms that TS values influenced TS reactivation, to the point of introducing incorrect mappings from other TS.

Lastly, we tested the influence of TS values on the creation of new TS, hypothesizing that higher-valued TS would influence TS creation more than lower-valued TS. This should be evident in the generalization phase of our task, in that participants would apply mappings from higher-valued TS more often than mappings from lower-valued TS. We found that participants chose actions according to TS0 (largest value), TS1 (intermediate value), and TS2 (lowest value) in an average of 30.1%, 23.9%, and 14.1% of trials, respectively, compared to chance levels of 3/12=25%, 3/12=25%, and 2/12=16.7% (Fig. 4). Participants chose actions that were correct in more than one TS in 22.4% of trials (chance 2/12=16.7%), and actions that were not correct in any TS in 9.4% of trials (chance 2/12=16.7%).

Table 4: Logistic mixed-effects regression predicting intrusion errors from stimulus values and TS values.

Task phase	Predictor	β	p
Initial learning	Stimulus value	0.01	0.09
	TS value	0.05	0.007
	Trial index	-0.10	<0.001
	Repetition	-0.18	<0.001
Refresher 1	Stimulus value	0.02	0.32
	TS value	0.01	0.68
	Trial index	-0.30	<0.001
	Repetition	-0.16	0.02
Refresher 2	Stimulus value	0.008	0.67
	TS value	0.13	0.003
	Trial index	-0.25	<0.001
	Repetition	-0.11	0.09
Hidden context	**Stimulus value**	0.02	0.035
	TS value	0.03	0.29
	Trial index	-0.24	<0.001
	Repetition	-0.12	0.05

To test for differences between TS, we analyzed the effect of TS values on the ratio of participant-selected to chance-expected choices, using linear regression. The effect of TS values was significant, controlling for two potential confounds, the values of individual stimulus-response mappings (low-level values), and participants' performance on each TS, a proxy for their confidence in the TS (table 5). In summary, these results suggest that action selection in the novel context was driven by previously-acquired TS, especially those of high value. Crucially, participants did not select items based on the values of individual alien-item mappings, or based on elevated confidence with certain TS.

Table 5: Linear mixed-effects regression predicting TS choices from stimulus values, TS values, and TS confidence.

Task phase	Predictor	β	p
Generalization phase	Stimulus value	-0.03	0.59
	TS value	0.14	0.045
	TS confidence	0.19	0.48

Discussion

We have shown evidence that supports a model of human learning about latent structure proposed by Collins & Frank (2013). In this model, complex Bayes-optimal reasoning is approximated by a simpler, biologically plausible architecture. Two RL loops are combined hierarchically that operate on state and action spaces at different levels of abstraction. In our task, participants learned different behavioral strategies (TS) in different contexts, and showed behavior consistent with predictions derived from this model.

Participants acquired coherent interdependent TS, replicating prior findings (Collins & Koechlin, 2012; Collins & Frank, 2013). Participants also acquired RL-like values at

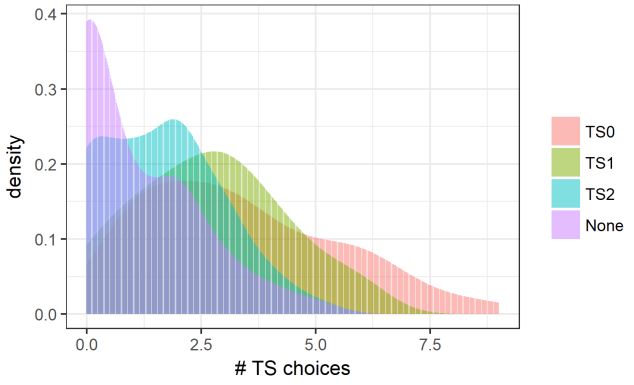

Figure 4: The distribution of TS choices in the generalization phase shows an effect of TS value.

different hierarchical levels in parallel, including the levels of stimuli (items), responses (aliens), and TS (contexts). Furthermore, TS values affected participants' behavior in multiple ways: participants learned faster and made fewer errors in higher-valued TS; their errors tended to reflect higher-valued TS; they preferred contexts that had been associated with higher-valued TS; and participants were more likely to generalize higher-valued TS to new contexts. Taken together, participants' behavior confirmed a sensitivity to TS values that is not predicted by many models of latent learning, such as the one proposed by Collins & Koechlin (2012). Our results are also inconsistent with flat RL models that lack context sensitivity or mechanisms to represent latent structure. The results are, however, compatible with our hypothesis that human learning is based on hierarchical RL mechanisms because in this model, TS selection is guided by TS values that are learned from reinforcement.

Nevertheless, other models might be able to account for our results as well. For example, a non-hierarchical neural network model including multiple dynamic scales has been shown to account for some aspects of behavioral and neural responses during TS learning (Bouchacourt, 2016), although it cannot capture all aspects of transfer. Non-hierarchical distributed models might account for our results if they are based on non-trivial mechanisms, such as joint learning of several context-stimulus-response pairs in conjunction with stochastic pattern completion of contexts. Models that are based on such clustering principles share their basic ideas with the notion of TS. Future work is necessary to arbitrate between potential models. We will implement our proposed as well as alternative models to quantify their competing predictions via simulations, and to allow for formal model comparison.

Another avenue for future research pertains to the neural structures that underlie structured, feedback-based learning. Our model is explicitly modeled to accord with the neural substrates that underlie RL and abstraction in the brain (Alexander et al., 1986). Future work needs to address these predictions specifically. Studies employing functional Magnetic Resonance Imaging (fMRI) will be necessary to assess whether the modeled processes are implemented in the predicted brain areas. Electroencephalography (EEG) could reveal whether the proposed mechanisms have their counterparts in patterns of brain activity, and shed light on the relationships between different processes.

Acknowledgments

We thank Lucy Whitmore and Sarah Master for helping with task design and data collection.

References

Alexander, G., DeLong, M., & Strick, P. (1986). Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex. *Annual Review of Neuroscience*, *9*(1), 357–381. doi: 10.1146/annurev.ne.09.030186.002041

Badre, D., & Frank, M. J. (2012). Mechanisms of Hierarchical Reinforcement Learning in CorticoStriatal Circuits 2: Evidence from fMRI. *Cerebral Cortex*, *22*(3), 527–536. doi: 10.1093/cercor/bhr117

Bouchacourt, F. (2016). *Hebbian mechanisms and temporal contiguity for unsupervised task-set learning*. phdthesis, Universit Pierre et Marie Curie - Paris VI.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. doi: 10.1037/a0030852

Collins, A. G., & Koechlin, E. (2012). Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLOS Biology*, *10*(3), e1001293. doi: 10.1371/journal.pbio.1001293

Daw, N., Gershman, S., Seymour, B., Dayan, P., & Dolan, R. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. doi: 10.1016/j.neuron.2011.02.027

Donoso, M., Collins, A. G., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486. doi: 10.1126/science.1252254

Frank, M. J., & Badre, D. (2012). Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 1: Computational Analysis. *Cerebral Cortex*, *22*(3), 509–526. doi: 10.1093/cercor/bhr114

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. doi: 10.1126/science.1102941

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. doi: 10.1146/annurev.neuro.24.1.167

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. doi: 10.1126/science.275.5306.1593

Sutton, R. S., & Barto, A. G. (2017). *Reinforcement Learning: An Introduction* (2nd ed.). Cambridge, MA; London, England: MIT Press.