What is a Choice in Reinforcement Learning?

Milena Rmus[1] and Anne G. E. Collins[1,2]

Department of Psychology[1], Helen Wills Neuroscience Institute[2],
University of California, Berkeley, Berkeley CA, 94704 USA

Corresponding author: milena_rmus@berkeley.edu

Abstract

In reinforcement learning (RL) experiments, participants learn to associate stimuli with rewarding responses. RL models capture such learning by estimating stimulus-response values. But what is a response? RL algorithms can model any response type, whether it is a basic motor action (e.g. pressing a key), or a more abstract, non-motor choice (e.g. selecting pizza at the restaurant). Are these different responses learned the same way? In this study, we examine differences between learning a rewarding association between (1) a stimulus and a motor action and (2) two stimuli. We show that learning differs between these two conditions, contrary to the common implicit assumption that response type does not matter. Specifically, participants were slower and less accurate in learning to select a rewarding stimulus. Using computational modeling, we show that the values of motor actions interfered with the values of stimulus responses, resulting in more incorrect choices in the latter condition.

Keywords: reinforcement learning; computational modeling; credit assignment; decision-making.

Introduction

The field of reinforcement learning (RL) provides a wealth of studies aiming to understand how individuals learn to make rewarding responses. Research in the field of RL has yielded great improvements in our understanding of the cognitive mechanisms that support the ability to select optimal choices, extending into more complex behaviors (e.g. planning: Daw et al., 2011; generalization: Niv et al., 2015). The RL framework also provides insights into developmental changes (Master et al., 2019) and clinical impairments (Gillan et al., 2016) in decision-making behavior.

Empirical RL research relies on variations of simple experimental designs in which participants learn rewarding associations between stimuli (i.e. a picture) and responses. In some experiments, the response is conceptualized as a motor action, such as a key press (Collins & Frank, 2012; Ratcliff and Frank, 2012). In other studies, the response instead consists of participants' selection of another visual stimulus (Daw et al., 2011; Foerde & Shohamy, 2014). Work on instrumental and classical conditioning has shown that selecting a motor action and approaching a goal may rely on different processes (Rescorla & Solomon, 1967). Furthermore, motor action and stimulus choice values are encoded differently in the brain (Luk & Wallis, 2013; Camille et al, 2011), and monkeys learning to select motor actions vs. stimuli behaved differently, and were affected differently by striatal lesions (Rothenhoefer et al., 2017). Despite this evidence that the response type is an important factor to consider when studying learning, human RL studies have not directly contrasted these types of responses. Consequently, conclusions from RL studies using one type of response (e.g. stimulus selection) are often implicitly assumed to generalize to the other type of response (e.g. motor action). Indeed, in most previous RL research, RL algorithms deployed to model behavior have treated these kinds of responses as equivalent (Daw et al., 2011; Collins, 2018). This propagates the implicit assumption that in the RL framework, learning of stimulus-stimulus (also referred to as goal-directed learning) and stimulus-motor action associations are identical processes that rely on the same mechanisms.

In this project, we investigate whether healthy young adults learn to select a response to a stimulus in the same way if the response is a motor action, or the selection of another (goal) stimulus in a reinforcement learning task. We use a novel experimental design and computational RL modeling to characterize the differences and similarities in learning. We provide evidence for a dissociation between two types of response learning, suggesting that 1) two response types are learned at different rates, and 2) the values of learned motor-action responses can impact the choice of the stimulus goal.

Methods

Participants

We recruited 82 participants (40 female, age mean (SD) = 20.5(1.93), age range = 18-30) from University of California, Berkeley participant pool. Participants received course credit as compensation for participating in the study. In accordance with the policy of the University of California, Berkeley Institutional Review Board, all participants provided a written informed consent before beginning the experiment. We excluded 20 participants due to insufficient learning performance as participants' average accuracy must exceed 0.60 in all three conditions, resulting in a total sample of 62 participants.

Experimental design

We developed a new task to directly compare how participants use reinforcement to learn stimulus-stimulus vs. stimulus-action associations. At the beginning of the experiment, participants received detailed task instructions and trial examples. Participants were told that on each trial, they would see one of the 6 cards from a card set (the stimuli), along with the 3 card boxes (responses). All boxes had distinct colors (red, green, blue), which we henceforth refer to as labels, and were placed in a left, middle, and right position on a line (Figure 1). The participants were asked to sort the cards into the boxes, based on different sorting rules. To sort a card, participants chose a box by pressing one of the three keys on the keyboard with the index, middle, and ring finger of their dominant hand, mapping motor actions onto the box positions (Figure 1). Following their response, participants received truthful feedback (+1 if they selected the correct box, 0 if they did not), before proceeding to the next trial.

We divided the task into three conditions based on different sorting rules, and counterbalanced the order of blocks to minimize the effect of block order on performance. In the label condition, participants were instructed to sort the cards based on the box label only. In other words, the correct box for each card was consistently defined by its label (Figure 1A), irrespective of its position on any given trial. In the position condition, participants were instructed to sort the cards only by the box position (Figure 1B). In this case, the correct box for each card was defined by the box position irrespective of its label on any given trial. In the position control condition (Figure 1C), the boxes were not tagged with labels, and participants could only sort cards by position. This condition allowed us to asses a baseline performance for the position condition, which was designed to visually match the label condition, with only one type of response possible.

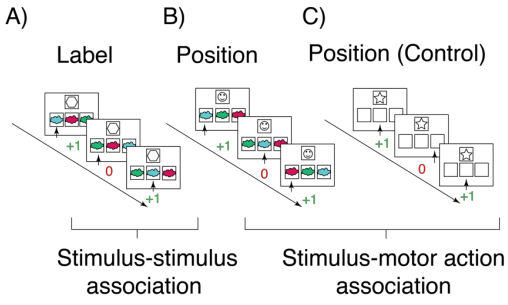

Figure 1. RL task with 3 conditions. In the first condition, participants learned an association between two stimuli (card and a label). In the two remaining conditions, participants learned an association between a stimulus and a motor action (card and a left/middle/right key press).

In each block, participants had to sort 6 cards into 3 boxes. The stimuli (cards) were different in each block, but the three box labels were the same across all blocks, except in the control condition where boxes were not labeled. Each card was presented 15 times for a total of 90 trials per block. We controlled the order of where the labels were placed on each trial, such that in the position condition, each label was shown on the correct box an equal number of times. In the label condition, the correct label appeared in each position an equal number of times. This order was counterbalanced, with equal distribution of unique label position pairing across different stages of the task.

On each trial, participants first saw the boxes for 1 second, with a fixation cross placed at the center of the screen. After 1 second, the card for the current trial replaced the fixation cross, and participants were allowed to make a response. Participants had 1 second to make their choice, after which they were provided with a 1s-feedback and a 1s inter-trial-interval before proceeding to the next trial. This trial timing was designed to alleviate a potential difficulty confound between the position and label conditions: it allowed participants to first identify where each box label was, as correct label selection was more demanding with label positions differing on each trial, before selecting a box in response to the stimuli.

We developed the different conditions (label/position) to elicit different response learning processes. Specifically, in the position condition, participants learned an association between a stimulus and a correct motor action (a card and a left, right or middle key press). In the label condition, participants learned an association between two stimuli (a card and one of the labels). We assessed the dissociation between the two types of response learning by addressing the following questions using model-independent analyses and computational modeling:

1) Are stimulus-action and stimulus-stimulus associations acquired differently, both qualitatively and quantitatively?

2) If so, what are the computational processes that drive the differences?

Results

We first plotted participants' learning curve (accuracy as a function of stimulus iteration) in all conditions. As expected, increased exposure to the cards and the subsequent truthful feedback increased participants' accuracy of choosing correct boxes (Figure 2). Repeated measures one-way ANOVA revealed an effect of condition on overall accuracy ($F(2,61) = 97.7$, $p = 4.5e\text{-}26$). Next, we sought to test which of the individual conditions differed significantly. The control and position condition, which both required learning of stimulus-motor action associations, were non-distinguishable (paired t-test: $t(61)=1.61$, $p=.11$), indicating that the alternating labels in the position condition did not impact performance. In contrast, participants' performance in the label condition was significantly worse than in the other conditions (paired t-test: position: $t(61) = 11.1$, $p = 3.8e\text{-}16$; control: $t(61) = 12.9$, $p = 5.4e\text{-}19$).

Next, we aimed to identify the mechanisms driving the decrease in accuracy in the label condition. Specifically, since label condition was more demanding due to labels changing positions on each trial, an uncompelling explanation for the respective drop in accuracy would attribute the higher frequency of errors to random slips in choices. This hypothesis would predict uniform, random error-patterns. Alternatively, we hypothesized that motor-action values could interfere with label values, suggesting an incorrect credit assignment to motor actions in the label condition, and thus incorrectly bias the choice of the stimulus response.

To test our hypothesis, we analyzed error types. We first computed a card-dependent reward history associated with each box label and each box position. Specifically, on each trial where participants received positive feedback, we incremented the cumulative reward history associated with the position and the label of the chosen box for the given card. Next, we used the label and position reward history to more carefully interpret choices on incorrect trials. We asked whether, out of the two possible incorrect choices, participants were more likely to make the one that had the highest past label reward history in the position condition, and the highest past position reward history in the label condition (henceforth referred to as interference errors). If the decreased performance in the label condition was due to noise, there should be no specific pattern in the errors. However, if it was due to interference, there should be more errors driven by position value.

We found that the proportion of interference errors was significantly greater than chance (0.5) in the label-sorting condition ($t(61) = 2.54$, $p = .01$), but not in position-sorting condition ($t(61) = .13$, $p = .89$; Figure 3A). The proportion of interference errors in the label condition was also significantly greater than that in the position condition ($t(61) = 2.69$, $p = .008$). To further confirm that the interference effect was not driven by participants mistakenly transferring the previous block's strategy, we ran a mixed-effects general linear model predicting accuracy as a function of current block condition and previous block condition. The results confirmed that accuracy was explained by current, but not previous block condition ($p = 2.22e-14$; $p=0.45$ respectively).

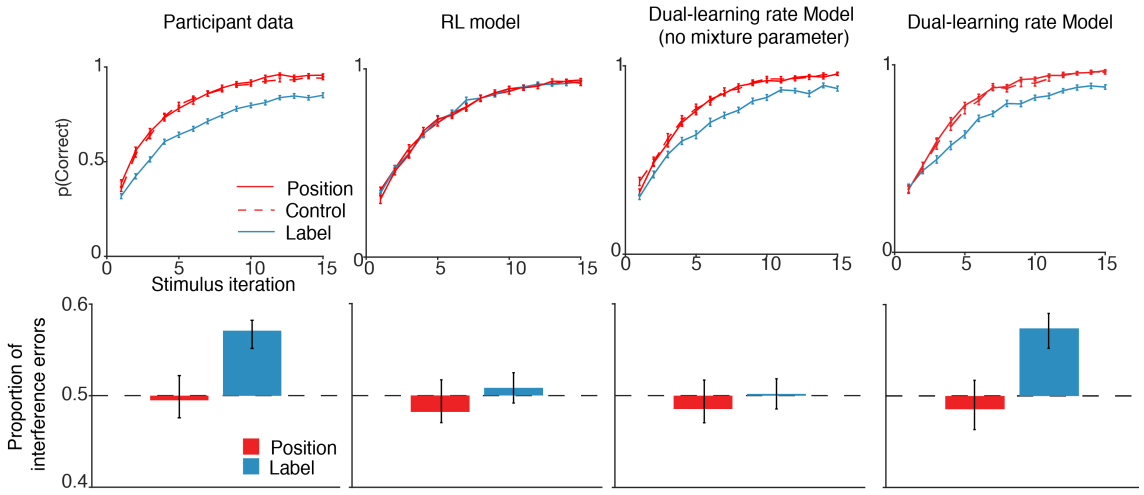

Figure 2. Learning curves and proportion of interference error types from participant data and simulations from the following models: RL, dual learning rate RL with no mixture parameter, and dual learning rate RL with mixture parameter. Baseline RL does not capture observed behavioral patterns. Dual learning-rate RL with no mixture parameter captures the difference in the learning curves, but not the interference errors. Only the dual rate model with possible deviation from correct policy captures both the difference in learning curves and the interference effect observed in behavioral data.

Next, we utilized the card-dependent label and position reward history of each box to examine whether response times also reflect interference in the choice process. First, we computed a trial-by-trial cumulative card-dependent reward history associated with positions and labels separately (Figure 3). Next, on each trial, we calculated the card-depended reward history difference (RHD) for both labels and positions. The RHD represented the difference between the reward history of the chosen box, and reward history sum of the non-chosen boxes. The RHD scaled with accuracy in the relevant condition. For instance, in the position condition, the correct box position was rewarded more frequently if participants were more accurate in the past. In contrast, incorrect box positions were never rewarded. Consequently, this led to a greater discrepancy between the cumulative reward history of the correct box position, and the cumulative reward history of the incorrect box positions. Given that participants' responses got faster with more learning, we hypothesized that greater RHD would predict faster response times only in correct dimension (i.e. the label RHD in label condition, and position RHD in position condition). On the other hand, we predicted the incorrect dimension (e.g. the label RHD in position block and vice versa) would have no effect on response times, unless there was an interference

effect. We performed a linear mixed effects model analysis, predicting log-transformed RTs on correct trials using position and label RHD. We controlled for the trial number, to ensure that the changes in the RTs were not simply driven by the practice effects or related factors associated with trial advancement.

Our results showed that participants' shorter RTs in the position and label conditions were indeed associated with higher respective RHD (label: β = -.04, p = 5.1e-19; position: β = -.06, p = 3.6e-21). The label RHD had no effect on RTs in the position condition (β= -.004, p = 0.55), supporting the conclusion that there was no interference of label values with the position choice. On the other hand, we found the opposite effect of position RHD in the label condition (β = .034, p = .001). Furthermore, subject level estimates of the incorrect factor RHD were significantly greater in label relative to position condition. (paired t-test: t(61) = 3.87, p = 2.6e-04). In other words, participant responses in label condition blocks were longer when position RHD was high (Figure 3B). This result complemented the error-type results, revealing an interference of motor-action values with learning of stimulus-stimulus associations. The asymmetry of the interference effect further implied that acquisition of stimulus-stimulus and stimulus-motor action associations in pursuit of rewards are not equivalent.

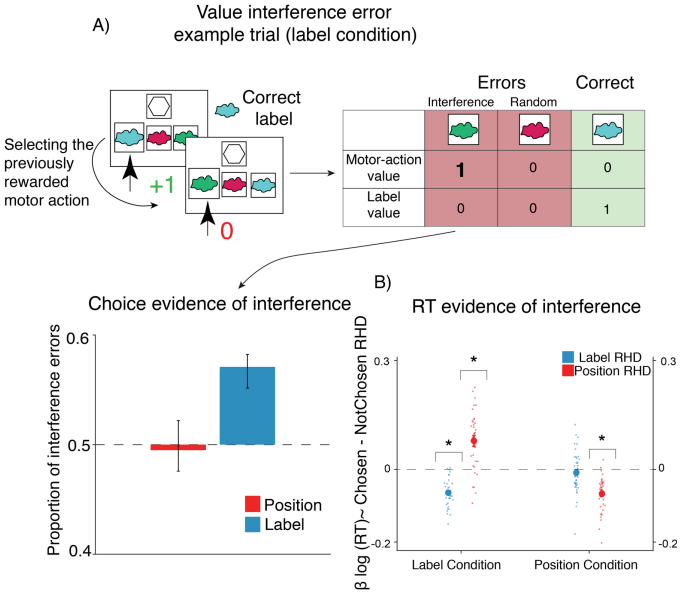

Figure 3. A) Example of the interference error trial in label condition. Participants track the values of both labels and conditions in parallel. On trials where evidence is accumulated in favor of a position not matching the location of the correct label, participants are more likely to select the response matching the box with high position value when incorrect. **B)** On correct trials participants tend to be slower when position and label values compete, which argues against the speed-accuracy tradeoff. This interference effect of incorrect dimension (the motor action) is specific to the label condition.

Model

Both error-type and RT analyses suggested that learning stimulus-motor action associations interferes with learning stimulus-stimulus associations: Participants did not fully segregate choice strategies across relevant conditions. We next sought to support these analyses with computational modeling in an effort to pinpoint the mechanisms of the interference effect. Was mixing of choice strategies essential for capturing data properties beyond the accuracy difference (i.e. exact patterns of errors)? Furthermore, did other mechanisms (i.e. the rate of learning or forgetting) differ between the 2 conditions, driving observed performance differences?

To answer these questions, we developed a reinforcement learning (RL) model of learning behavior in this task. Our model assumes that (1) both values of positions and labels are learned and updated individually, and (2) there is a mixture of choice strategies, such that each choice may reflect a contribution of both the position and the label value, allowing for potential interference when only one value is relevant (Figure 3).

The family of RL models we considered extended a classic model-free RL (Sutton and Barto, 1998; Schultz, Dayan & Montague, 1997) with two main parameters: learning rate and softmax inverse temperature. We integrated additional processes to parameterize different aspects of behavior that basic RL alone does not capture.

Our model assumed feedback-dependent value-learning in both conditions, in that for each card *c* the expected reward of the correct labels $Q^L(c,l)$ and positions $Q^P(c,p)$ was incrementally updated based on the outcome of each trial. The reward history for both the label and the position of the selected box was updated as a function of prediction error between expected and the observed outcome at trial t:

$$Q^P{}_{t+1}(c,p) = Q^P{}_t(c,p) + \alpha \times \delta P$$
$$Q^L{}_{t+1}(c,l) = Q^L{}_t(c,l) + \alpha \times \delta L$$

where δ was the dimension-specific reward prediction error, formalized as $\delta_t = r_t - Q(c,\text{response})$, and α was the learning rate. Although value updating was identical for labels and positions, we assumed separate reward prediction errors for label and position.

Choices of position/label with greater Q-values were selected with a greater likelihood, as a function of softmax choice policy:

$$P(position|c) = \frac{\exp(\beta * Q^P(c,\text{position}))}{\Sigma_p \exp(\beta * Q^P(c,p))}$$

$$P(label|c) = \frac{\exp(\beta * Q^L(c,\text{label}))}{\Sigma_l \exp(\beta * Q^L(c,l))}$$

where β was the inverse temperature, which controlled the stochasticity in the choice policy, based on the differences between the values of each response. Importantly, we assumed that the final policy was a mixture of two choice

strategies: (1) choosing the box with the highest position and (2) highest label value at the policy level, such that:

$P(position|c, pos.\ block) = \rho_P * p(position) + (1- \rho_P) * p(label)$
$P(label|c, lab.\ block) = \rho_L * p(label) + (1- \rho_L) * p(position)$

In both conditions, higher values of the ρ parameter indicated that the choices were influenced by the value of the relevant dimension (i.e. values of labels in label condition and values of positions in position condition). $\rho < 1$ indicated deviation from the correct policy, and an influence of the incorrect value dimension (i.e. position values in label condition). The policy in the control blocks was identical to the one in position blocks.

In addition to the stochasticity of the choice which the softmax allowed, the undirected noise parameter e allowed the model to capture value-independent random slips in choices (Nassar & Frank, 2016). We defined a new policy by incorporating the undirected noise into the choice process:

$$P' = (1-\varepsilon)*p + \varepsilon * \frac{1}{nC}$$

where nC was the number of choices, $\frac{1}{nC}$ was the uniform random policy, and ε was the noise parameter ($0< \varepsilon <1$). Higher value of ε indicated higher likelihood of random lapses. We also implemented forgetting by allowing the Q-values of positions and labels to decay on each trial:

$$Q^P_{t+1} = Q^P_t + d*(Q^{P0}-Q^P_t),$$
$$Q^L_{t+1} = Q^L_t + d*(Q^{L0}-Q^L_t),$$

Where d ($0<d<1$) was the decay parameter. Higher d values indicated faster forgetting.

Prior work in similar tasks (Collins, 2018; Christakou et al., 2013) has shown that individuals tend to learn less from negative than positive feedback. To capture individuals' propensity to neglect negative feedback, we also integrated a learning bias parameter, such that for negative prediction errors, the learning rate is reduced to α *(learning bias).

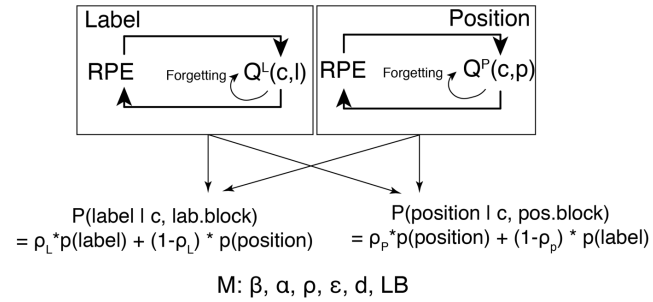

Figure 4. Model schematic. The model assumes different learning and value updating for labels and positions. The box choice is assumed to be determined by a mixture of position and label values.

We used the Matlab optimization function fmincon (the Mathworks Inc., Natick, Massachusetts, USA) to fit parameters with 20 randomly chosen starting points to reduce the likelihood of finding a local rather than global minimum. We fit our models to each participant's data individually using maximum likelihood estimation method. We fit all the parameters, except β and ρ_P, with a lower bound = 0 and upper bound =1. Following previous work (Collins, 2018; Master et. al, 2019), we observed that fixing β =100 improved parameter recovery and estimation. Note that leaving β free or condition-dependent did *not* improve the model fit. Following behavioral results that showed 1) no performance difference between control and position conditions, and 2) no interference of label value in position conditions (Figure 2), we also fixed ρ_P to 1. We confirmed with model comparison that leaving ρ_P a free parameter did not improve model fit (see Modeling results).

Model comparison

We repeated the fitting and the simulation procedure for the models listed in the Table 1. We aimed to test whether placing different constraints, such as allowing only one or no distinct parameters for label and position learning, would enable us to capture the behavior better. We compared the tested models using the Akaike Information Criterion (AIC) which penalizes model complexity. For each of the models of interest we simulated data, then fit all of the models to the simulated data. We were able to recover the ground truth (simulating model) via model comparison with AIC (Figure 5), confirming that AIC is appropriate for model comparison in this context (Wilson & Collins, 2019).

Modes	Learning rate(α)	Decay(d)	Noise(ε)	Mixture (ρ)
M1	1	1	1	1(ρ_L)
M2	1	1	2	1(ρ_L)
M3	2	1	1	0
M4	2	1	1	1(ρ_L)

Table 1. List of the model variations we compared. For each column, the values of 1 and 2 indicate whether we used a single parameter for both label and position learning, or two distinct parameters respectively. * ρ_L = mixture parameter which weighs the value of labels; 2 learning rates = distinct learning rates for learning position and label values.

Specific models

We first verified that including undirected noise, forgetting, and learning bias improved our ability to capture behavior relative to 2-parameter RL. Next, we verified that including a free ρ_L parameter improved the fit, while leaving β and ρ_F free did not. Thus, our baseline model M1 included 5 free parameters (α , ρ_L, ε, d, learning bias). This baseline model treats the different conditions as identical by utilizing the same set of parameters for all conditions, except for the label-condition-specific ρ_L parameter.

We compared this baseline model to a family of models that allowed for more graded differentiation between the response learning processes in the two conditions (control condition was treated as identical to position condition). Systematically varying the structure and complexity of the compared models allowed us to identify the best fitting model, and in doing so isolate the cognitive mechanisms that most likely drive differences in stimulus-stimulus vs. stimulus-action learning. In particular, we tested the models with different combinations of condition-dependent learning rate, decay, learning bias and undirected noise. Here, we focus solely on a few models that enabled us to test specific theoretical predictions regarding the difference in choice processes in stimulus-stimulus and stimulus-action associations (Table 1). Specifically, dual noise RL model M2 tests whether the observed condition dissociation in learning can be explained by stimulus-value learning being a noisier/more difficult process. Dual learning rate RL with fixed mixture parameter ρ_L=1 M3 tests whether the empirical dissociation can be captured solely by different learning rates. Last, model M4 (dual learning rate RL with free mixture parameter ρ_L) tests whether a mixture policy is necessary to capture the full behavioral pattern, including error types. Other models did not sufficiently account for and fit the data, thus we omitted them from further discussion.

Model validation

To validate the models' fit to the data (Palminteri et al., 2017, Wilson & Collins, 2020), we tested whether they captured key qualitative features of behavior with high fidelity. Specifically, for each participant, we simulated the models using individually fit parameters 100 times and averaged the simulations' performance to capture the model's predicted behavior for that participant (Figure 2).

Modeling results

The AIC comparison revealed that the model with two learning rates, single decay and noise parameters and a free ρ_L mixture parameter (M4, Table 1) had the best fit relative to other models (Figure 5A). Model simulations revealed that this model captured the critical features of the participants' behavior (Figure 2). Specifically, simulated accuracy of M4 in the label condition was lower than simulated accuracy in the position and control conditions. Simulations captured both the lower accuracy in the early learning stages (averaged over first several stimulus iterations) as well as the asymptotic accuracy (accuracy over later stimulus iterations). By contrast, a classic RL model could not capture observed behavior (Figure 2).

The validation and model comparison results, therefore, supported the conclusion that the dissociation between stimulus-stimulus and stimulus-motor action association learning was primarily driven by the difference in learning rates, rather than decay or the rate of random slips in actions.

To test the necessity of the mixture policy in label blocks to capture condition effects in behavior, we fit the dual learning rate model without the mixture parameter (M3, Table 1). We found that this model fails to produce the observed interference errors, suggesting that the mixture parameter is essential to capture the contribution of different values to the choice process (Figure 2). Finally, to quantify the asymmetry in the interference effect, we fit an addition model that included the same parameters as the winning model (dual learning rate and mixture parameter) with an additional free ρ_P parameter. This model did not improve the fit, confirming that a fixed ρ_P =1, ρ_L <1 captured the data well.

We next sought to assess the differences in condition-dependent parameters. In the winning model M4, condition comparison revealed that the learning rate in the position condition was significantly greater than the learning rate in the label condition (sign test p = 7e-10; Figure 5c).

The computational modeling approach allowed us to decouple the mechanisms contributing to learning correct stimulus, and correct motor-action responses. The basic RL model with a single set of parameters for both types of response learning failed to capture the data in both conditions, thus suggesting different learning processes for stimulus-stimulus and stimulus-motor action associations. Since the model comparison favored the model with (1) dual learning rates and (2) different policy mixture parameters, we concluded that the underlying learning mechanisms of stimulus-stimulus and stimulus-motor action associations are not equivalent, as commonly assumed. Furthermore, given that this model provided the best fit, we reasoned that while different decay, learning bias, and undirected noise parameters could contribute to capturing the behavioral features, they were not essential for explaining the differences between these conditions. Thus, combining model-independent error-type and RT analyses with modeling allowed us to confirm that behavioral differences in the two conditions cannot be explained by a discrepancy in condition difficulty and performance noise.

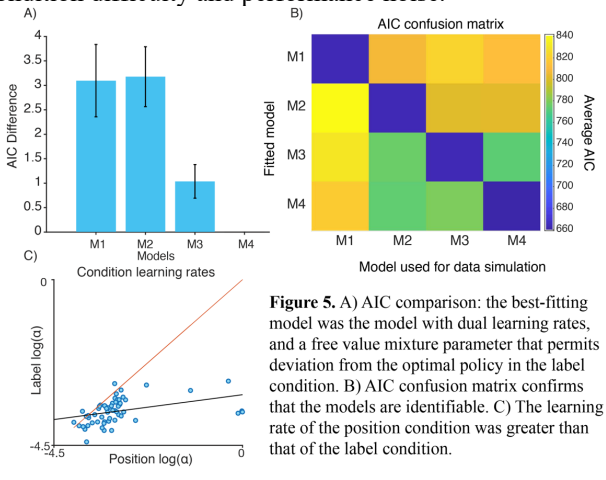

Figure 5. A) AIC comparison: the best-fitting model was the model with dual learning rates, and a free value mixture parameter that permits deviation from the optimal policy in the label condition. B) AIC confusion matrix confirms that the models are identifiable. C) The learning rate of the position condition was greater than that of the label condition.

Discussion

We combined a novel experimental design and computational modeling approach to test the equivalence of stimulus-stimulus and stimulus-motor action associations in

the RL framework. Consistent with previous work on instrumental learning, we presented results that challenge the homogeneity of response definition in RL. Specifically, we showed that learning processes with different learning rates underlie stimulus and motor action response-learning. This contradicts the implicit assumption that all response types are equivalent, inviting for caution in future reinforcement learning studies and modeling practice.

Our current design prohibits us from testing whether the stimulus-stimulus associations are always more 'suboptimal' and susceptible to interference (i.e. from other, less correct stimulus responses) relative to stimulus-motor action associations. Future work is also required to disambiguate mechanisms of interference – for instance, whether the interference effect is driven solely at the policy level, or also incorrect stimulus and motor-action value updates. In addition, while we attempted to dissociate between stimulus and motor responses, we cannot rule out the possibility that the position sorting condition represents another form of stimulus response (e.g. based on spatial position).

Our results showed that participants track the value of motor actions even when they are irrelevant (in the label condition), and that irrelevant action values influence their choices. This was revealed by the interference errors, where the value of the irrelevant position dimension influenced which error was made, and was captured by a mixture parameter in the RL model. This pattern in errors also ruled out the possibility that condition effects were due to a difficulty confound, an explanation also ruled out by the worse fit of a model including multiple noise parameters. Thus, our results suggest that the dissociable strategies based on different sets of values compete during the choice process. This highlights the possibility of parallel RL circuits in the brain contributing jointly to decision-making, and the importance of clearly defining response types in RL studies.

References

Camille, N., Tsuchida, A., & Fellows, L. K. (2011). Double Dissociation of Stimulus-Value and Action-Value Learning in Humans with Orbitofrontal or Anterior Cingulate Cortex Damage. Journal of Neuroscience, 31(42), 15048–15052.

Christakou, A., Gershman, S.J., Niv, Y., Simmons, A., Brammer, M., Rubia, K. (2013). Neural and psychological maturation of decision-making in adolescence and young adulthood. Journal of Cognitive Neuroscience, 25, 1807–1823.

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

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. European Journal of Neuroscience, 35(7), 1024–1035.

Daw, N.D. (2011) Trial-by-trial data analysis using computational models, in: Delgado M., Phelps E.A., and Robbins T.W. (eds.) Decision Making, Affect, and Learning, Attention and Performance XXIII, Oxford University Press.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron. 69:1204 1215.

Foerde, K., & Shohamy, D. (2011). Feedback Timing Modulates Brain Systems for Learning in Humans. Journal of Neuroscience, 31(37), 13157–13167.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. ELife,5 .

Luk, C.-H., & Wallis, J. D. (2013). Choice Coding in Frontal Cortex during Stimulus-Guided or Action-Guided Decision-Making. Journal of Neuroscience, 33(5), 1864 1871.

Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. (2019). Distentangling the systems contributing to changes in learning during adolescence.

Nassar, M. R., & Frank, M. J. (2016). Taming the beast: Extracting generalizable knowledge from computational models of cognition. Current Opinion in Behavioral Sciences, 11, 49–54.

Niv, Y., Daniel R., Geana, A., Gershman, S.J., Leong Y.C., Radulescu, A., Wilson, R.C.. Reinforcement learning in multidimensional environments relies on attention mechanisms. The Journal of Neuroscience : the Official Journal of the Society For Neuroscience. 35: 8145-57.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. Trends in cognitive sciences, 21(6), 425–433.

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.

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review*, *74*(3), 151–182.

Rothenhoefer, K. M., Costa, V. D., Bartolo, R., Vicario-Feliciano, R., Murray, E. A., & Averbeck, B. B. (2017). Effects of Ventral Striatum Lesions on Stimulus-Based versus Action-Based Reinforcement Learning. *Journal of Neuroscience*, *37*(29), 6902–6914.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. Science, 275(5306)., 1593–1599.

Sutton, R.S., & Barto, A.G. (1998). Reinforcement learning: An introduction. MIT Press.

Wilson, R. C., & Collins, A. (2019). Ten simple rules for the computational modeling of behavioral data. Elife.