Plucking a string or playing a G? Degree of choice abstraction impacts human reinforcement learning

Milena Rmus[1], Amy Zou[1], Anne G.E. Collins[1 2]
University of California, Berkeley
[1]Department of Psychology
[2]Helen Wills Neuroscience Institute
Berkeley, California 94720, USA

August 25, 2021

Abstract

In reinforcement learning (RL) experiments, participants learn to make rewarding choices in response to different stimuli; RL models use outcomes to estimate stimulus-response values which change incrementally. RL models consider any response type indiscriminately, ranging from less abstract choices (e.g. pressing a key with the index finger), to more abstract choices that can be executed in a number of ways (e.g. getting dinner at the restaurant). But does the learning process vary as a function of how abstract the choices are? In Experiment 1, we show that choice abstraction impacts learning: participants were slower and less accurate in learning to select a more abstract choice. Using computational modeling, we show that two mechanisms contribute to this. First, the values of motor actions interfered with the values of more abstract responses, resulting in more incorrect choices; second, information integration for relevant abstract choices was slower. In Experiment 2, we replicate the findings from Experiment 1, and further extend the results by investigating whether slowed learning is attributable to working memory (WM) or RL contributions. We find that the impairment

in more abstract/flexible choices is driven primarily by a weaker contribution of WM. We conclude that defining a more abstract choice space used by multiple learning systems recruits limited executive resources, limiting how much such processes then contribute to fast learning.

1 Introduction

The ability to learn rewarding choices from non-rewarding ones lies at the core of successful goal-directed behavior. But what counts as a choice? When a child tries the pink yogurt in the left cup, and the white yogurt in the right cup, then prefers the right cup, what choice should they credit this rewarding outcome to? In their next decision, should they repeat their previously rewarding reach to the yogurt on the right, independently of its color, or should they figure out where the white yogurt is before reaching for it? Making the more abstract choice (selecting the type of yogurt) still involves a subsequent motor choice, but in this case, the motor choice should not be credited. Knowing the relevant choice to assign credit to is essential when learning. How does the level of abstraction of our choices impact how we learn?

The framework of reinforcement learning (RL) is a highly successful theoretical framework to study reward-based learning and credit assignment (Sutton & Barto, 2018). However, RL as a computational model of cognition typically assumes a given action space defined by the modeler, which provides the relevant dimensions of the choice space (i.e. either the yogurt color or the cup position) - there is no ambiguity in what choices are, and the nature of the choice space does not matter (Rmus, McDougle, & Collins, 2020). As such, RL experiments in psychology tend to not consider the type of choices (a single motor-action such as pressing a key with the index finger; (A. G. E. Collins, Ciullo, Frank, & Badre, 2017; Tai, Lee, Benavidez, Bonci, & Wilbrecht, 2012), or the selection of a goal stimulus (Foerde & Shohamy, 2011; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Frank, Moustafa, Haughey, Curran, & Hutchison, 2007)) as important, and researchers use the same models and generalize findings across choice types. Recent research has shed some light into how participants might identify relevant dimensions of the state and choice space (Niv, 2019; Farashahi, Rowe, Aslami, Lee, & Soltani, 2017); however, this research does not address how learning occurs

when the choice space is known but enables multiple degrees of abstraction, such as in our yogurt example.

Examining learning of responses with varying degrees of abstraction is important, however, as most of our choices in everyday life can be represented at multiple hierarchical levels (e.g. have breakfast; have yogurt; have pink yogurt; reach for the yogurt on the right side, etc.). Furthermore, different types of choices may recruit different cognitive/neural mechanisms (Rescorla & Solomon, 1967). For example, previous animal models of decision-making suggest that the orbitofrontal cortex and the anterior cingular cortex index choice outcomes for more abstract stimulus choices and motor action choices respectively (Luk & Wallis, 2013). Ventral striatum lesions in monkeys impaired learning to choose rewarding stimuli (more abstract), but not rewarding actions (less abstract) (Rothenhoefer et al., 2017). In humans, recent behavioral evidence suggests that the credit assignment process is what differentiates learning more abstract from less abstract rewarding choices (McDougle et al., 2016), and that there might be a hierarchical gradation of choices in terms of credit assignment. Specifically, while people are capable of learning the value of both abstract rule choices and concrete action choices in parallel (Eckstein, Starr, & Bunge, 2019; Eckstein & Collins, 2019; Ballard, Miller, Piantadosi, Goodman, & McClure, 2018), they also seem to assign credit by default to more concrete actions when making abstract choices realized through motor actions (Shahar et al., 2019)).

The brain relies on multiple neuro-cognitive systems for decision-making; whether choice format impacts learning similarly across systems remains unexplored. Specifically, while RL models provide a useful formalism of learning, they do not easily relate to underlying processes. Indeed, RL models are known to summarize multiple processes that contribute to learning jointly (Eckstein, Wilbrecht, & Collins, 2021), such as the brain's RL mechanism, but also episodic memory (Wimmer & Shohamy, 2012; Bornstein & Daw, 2013; Vikbladh et al., 2019; Bornstein, Khaw, Shohamy, & Daw, 2017; Poldrack et al., 2001), or executive functions (Rmus et al., 2020; A. G. E. Collins & Frank, 2012). Here we focus on working memory (WM), which has also been shown to contribute to learning alongside RL (A. G. E. Collins & Frank, 2012, 2018; A. G. E. Collins et al., 2017). If choice type matters for learning, does it matter equally for each cognitive system that contributes to learning, or differently so?

In summary, there is a two-fold gap in our understanding of how choice format impacts learning. First, does choice type impact learning, and if so,

what are the computational mechanisms that differ in learning reward associations to more or less abstract choices when these are known and directly compared? Second, are the differences rooted in the brain's RL system, WM, or both? To address this gap, we designed a task that directly compares learning to make choices with different levels of abstraction, when there is no ambiguity about which choices are relevant to the learning problem. In our task, the less abstract choices are conceptualized as motor actions, and more abstract choices are conceptualized as the selection of stimulus goals that constrain a downstream motor action. In a second experiment, we manipulated learning load to separately identify WM and RL contributions to learning, and investigated with computational modeling how choice matters in both systems.

Our results across two experiments suggest that choice type impacted learning overall, resulting in slower learning when choices were more abstract. This was in part driven by an incorrect, asymmetric credit assignment to less abstract choices when they were irrelevant. Furthermore, WM (rather than RL) mechanisms seemed to drive the deficits in performance in the more abstract choice condition, indicating that defining a more abstract action space, shared by multiple choice systems, recruited limited executive resources.

2 Methods

2.1 Participants

2.1.1 Experiment 1

Our sample for experiment 1 consisted of 82 participants (40 female, age mean (SD) = 20.5(1.93), age range = 18-30), recruited from the University of California, Berkeley Psychology Department's Research Participation Program (RPP). In accordance with the UC Berkeley Institutional Review Board policy, participants provided written informed consent before taking part in the study. They received course credit for their participation. To ensure that the participants included in analyses were engaged with the task, we set up an exclusion criterion of 0.60 or greater average accuracy across all task conditions. We excluded 20 participants based on this criterion, resulting in a total sample of 62 participants for the reported analyses.

2.1.2 Experiment 2

For the second experiment, we recruited 75 participants (54 female, 1 preferred not to answer; age mean (SD) = 20.34(2.4), Age-range=18-34) from the University of California, Berkeley RPP. Participants completed the experiment online (de Leeuw, 2015), and received course credit for their participation. Using the same exclusion criteria as previous experiment, we excluded 18 participants, resulting in the total sample of 57 participants.

2.2 Experimental protocol

2.2.1 Experiment 1

Learning phase structure. Participants were instructed that they would be playing a card sorting game, and that on each trial they would sort a card into one of three boxes. Their goal was to use reward feedback to learn which box to sort each card in. The boxes were labeled with 3 different colors (green, blue and red), and participants chose one of the boxes by pressing one of three contiguous keyboard keys (corresponding to the box position) with their index, middle and ring finger. Importantly, the color of the boxes changed positions on different trials (i.e. the blue box could appear on the right side on trial n, and in the middle on trial n+1). Participants received deterministic feedback after each selection (+1 if they selected the correct box for the current card, 0 otherwise).

Before the experiment, participants read detailed instructions and practiced each task condition. The task then consisted of 8 blocks, divided into three conditions. Each of the three conditions was defined by its distinct sorting rule. In the label condition, the correct box for a given card was defined deterministically by the box's color label (Fig. 1A). For instance, if the blue box was a correct choice for the given card, participants were always supposed to select the blue box in response to that card, regardless of which respective key mapped onto the blue box on a given trial. In the position condition, the correct box was defined deterministically by the box's position (left/middle/right). For example, the correct response of a given card would always be achieved by pressing the leftmost key with the index finger, regardless of the box color occupying the left position (Fig. 1B). The sorting rule in the position control condition was identical to the sorting rule in the position condition, but the boxes were not tagged with color labels. This

condition allowed us to assess participants' baseline performance when only one response type (e.g. position, but not the label) was available. Importantly, participants were explicitly told the sorting rule (position or label) at the beginning of each block, in order avoid any performance variability that may arise as a function of rule inference and uncertainty.

Out of 8 blocks in total, 2 were control condition blocks, 3 were position condition, and 3 were label condition. Block order was pseudo-randomized: participants completed a control block first and last, while blocks 2-7 conditions were randomly chosen within subjects, but counterbalanced across subjects. In each block, participants learned how to sort 6 different cards; we used a different set of images to represent cards in each block. The boxes were labeled with the same 3 colors across all blocks, except the position control blocks, where the boxes were not labeled. Participants experienced 15 repetitions of each card, resulting in 90 trials per block; trial order was pseudo-randomized to ensure a uniform distribution of delays between repetitions of the same card in a block. We controlled for the card-dependent position-label combinations across trials. Specifically, each label occurred in each position an equal number of times (i.e. the blue label occurred 5 times on the left, right and middle box for each card). We also ensured that the position-label combinations were evenly distributed across the task (i.e. blue-middle combination did not occur only during the first quarter of block trials).

Single trial structure. On each trial, participants first saw the three boxes with their color labels underneath a fixation cross at the center of the screen. After 1 second, the card appeared in the center of the screen, replacing the fixation cross. Participants were allowed to press a key only once the card appeared, with a 1-second deadline. Following their response, participants received feedback (+1 or 0) that remained on the screen for 1 second, followed by a 1 second inter-trial interval (fixation cross). This trial structure allowed allowed participants to identify where each color label was positioned, thus minimizing a potential advantage of the position conditions, where participants did not need to know where colors were on a trial-by-trial basis in order to make a correct response.

We designed the label and position condition to engage choice processes with different degrees of abstraction. The position condition should capture the less abstract choice process in which the rewarding response is defined

by a single motor action. The label condition, on the other hand, captures a more abstract choice process in which the rewarding response (i.e. choice of the correct label) can be made by executing any of the three motor actions, depending on where the correct box label is positioned on the given trial.

Test phase. Following the training phase, participants performed an unrelated task (N-back, data not analyzed in this paper). Unbeknownst to them, there was a test phase following the delay introduced by the N-back task, where participants were asked whether two cards were sorted in the same box or different boxes in the training phase. Test phase results are beyond the scope of this paper and not reported here.

2.2.2 Experiment 2

The task design for Experiment 2 was the same as the task design for Experiment 1, with one important exception - we varied the number of cards per block between 2 and 5, for both position and label conditions. This manipulation has previously been shown to enable computational modeling to disentangle working memory and reinforcement learning processes (A. G. E. Collins & Frank, 2012). The order of blocks was counterbalanced across participants; participants completed either label or position blocks first, with the order of set sizes randomized for the first completed condition, and then repeated for the second. In addition, we removed the control condition, given that we previously observed no difference between position and control. Participants completed 4 blocks of position and label each, where each condition block had a different set size.

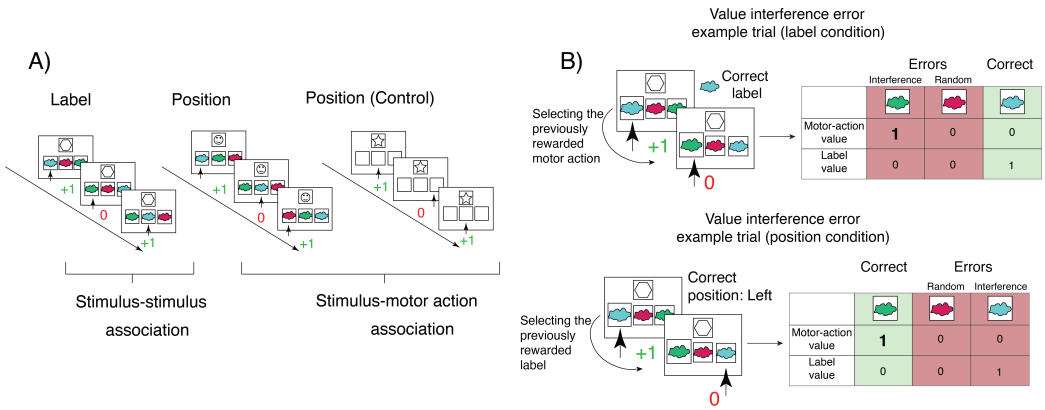


Figure 1: Experiment design. A) Participants played a card-sorting game with 3 different conditions: Label (learning which box color is correct for each card - more abstract choice), Position (learning which motor action/position is correct for each card - less abstract choice), Control (identical sorting rules as position condition, but without label-tagged boxes). B) We assumed that participants track card-dependent reward history for both positions and labels, and that both of these contribute to choice selection process, sometimes resulting in interference errors. Note that the card-dependent reward history is cumulative (tracked across all past trials during which the given card was presented, rather than only one-trial back), but for simplicity of illustration we only show 1-back trial in the panel B.

.

2.3 Analyses

2.3.1 Model-independent

We first examined participants' performance by evaluating their average accuracy and log-transformed response times (RT) as a function of stimulus iteration/accumulated evidence. These analyses provided us with a general insight into how the learning process evolved over time in each condition, using both the choice measure (accuracy) and time required to execute responses (RT).

Next, we sought to analyze participants' choices and reaction times as a function of how often each motor action and each label had been rewarded for each card. Specifically, we computed card-dependent cumulative reward

history (CRH) for both positions P and labels L on each trial, in each condition:

$$CRH_t^P(C,P) = \sum_{k=1}^{t} (r_k * 1(Card_t = C, Choice_t = P) \tag{1}$$

$$CRH_t^L(C,L) = \sum_{k=1}^{t} (r_k * 1(Card_t = C, Choice_t = L) \tag{2}$$

where r_k is the outcome at trial k in the block, and 1 is the indicator function taking value 1 if the card and position/label match C and P/L, and 0 otherwise. We used this metric to analyze how the integration of two value sources shaped choices when choice format was less/more abstract. In particular, in the example of the position condition, the position CRH for a card and its associated correct position indicated the past number of correct choices, while the CRH for other positions was 0. By contrast, in the same position condition, the label CRH for a card reflected how often each label had been rewarded due to this label being in the correct position. All label CRH values in the position condition were expected to be close to each other because label positions was counterbalanced, but slight differences due to past choice randomness could be predictive of biases in future choice. The opposite was true in the label condition.

To analyze how the value integration for each type of choice shaped decisions, we focused on the error trials and computed the proportion of errors driven by the other irrelevant choice dimension. We reasoned that if participants were lapsing randomly, any of the two possible errors should be equally likely. However, if participants experienced value interference, they should be more likely to select the error with the higher CRH in the irrelevant dimension. In the label condition, such an interference error would look like selecting the position/motor action that was rewarded on the previous trial, even though the correct label had switched positions since (Fig. 1B). In the position condition, an interference error would occur when participants selected the previously rewarded label that had switched positions, instead of the label corresponding to the same position/motor action that is always correct for the given card (Fig. 1B).

In addition to general diagnostics, we also ran a trial-by-trial analysis using a mixed effects general linear model. We used trial-by-trial reward history difference $RHD = CRH(chosen) - mean(CRH(unchosen))$ between chosen

and unchosen boxes, for both positions and labels, and tested whether this discrepancy modulated accuracy and RTs.

If participants implemented an optimal decision strategy, their accuracy and RTs should increase and decrease respectively with an increased RHD in the relevant choice dimension (i.e. label RHD in label condition, position RHD in position conditon). Alternatively, contribution of the irrelevant dimension RHD (i.e. position RHD in label condition or vice versa) would serve as evidence of value interference.

2.3.2 Computational modeling - Experiment 1

Reinforcement Learning (RL): In order to computationally quantify the differences in learning processes in these two conditions, we developed and fitted a set of reinforcement learning (RL) models to our data. Our baseline assumption, for all model configurations, was that participants track and update two independent sets of stimulus-action value tables, corresponding to the two possible choice spaces: a card-position value table, and a card-label value table. We also assumed that the choice selection may reflect a mixture of both the relevant and the irrelevant value table, such that interference errors occur when the value of irrelevant choice dimension (position/label) contributes to the choice process (Fig. 2A). We pinpointed the computational mechanisms of divergence between the learning processes in the two conditions by varying the extent to which the model allows for condition-dependent specificity/model-parameters.

The basic RL model assumes incremental learning through a simple delta rule (Sutton and Barto, 1998). Specifically, on each trial t, the values of labels $Q_L(c,l)$ and positions $Q_P(c,p)$ for the trial's card c and chosen labels and positions l and p are updated in proportion to the reward prediction error:

$$Q^P_{t+1}(c,p) = Q^P_t(c,p) + \alpha * (r - Q^P_t(c,p)) \tag{3}$$

$$Q^L_{t+1}(c,l) = Q^L_t(c,l) + \alpha * (r - Q^L_t(c,l)) \tag{4}$$

where α is the learning rate, and $r = 0,1$ is the outcome for incorrect and correct trials. Q-tables are initialized at $1/3$ at the start of each block to reflect initial reward expectation in the absence of information about new

cards. We transformed the values into choice probabilities using the softmax function:

$$P_p(position|card) = \frac{exp(\beta * Q_t^P(c,p))}{\sum_{i=1}^{3} exp(\beta * Q_t^P(c,p_i))}, \tag{5}$$

$$P_L(label|card) = \frac{exp(\beta * Q_t^L(c,l))}{\sum_{i=1}^{3} exp(\beta * Q_t^L(c,l_i))}, \tag{6}$$

selecting choices with higher Q value with higher probability. The softmax β is the inverse temperature parameter, which controls how deterministic the choice process is.

The overall choice policy is a mixture of both policies, determined by mixture parameters, ρ:

$$P(position_i|pos.block) = \rho_P * P_p(position_i) + (1-\rho_P) * P_L(label(position_i)), \tag{7}$$

$$P(label_i|lab.block) = \rho_L * P_L(label_i) + (1-\rho_L) * P_p(position(label_i)), \tag{8}$$

We further modified the RL algorithm in order to parameterize additional processes. For instance, individuals often make value-independent, random lapses in choice while doing the task. To capture this property of behavior, we derived a secondary policy by adding a random noise parameter in choice selection (Nassar & Frank, 2016):

$$P' = (1-\varepsilon) * P + \varepsilon * \frac{1}{n_A}, \tag{9}$$

where n_A is the total number of possible actions, $1/n_A$ is the uniform random policy, and ε is the noise parameter capturing degree of random lapses.

Prior work in similar tasks (Sugawara & Katahira, 2021; Frank et al., 2007; Gershman, 2015; Niv, Edlund, Dayan, & O'Doherty, 2012) has shown an asymmetry in learning based on positive/negative feedback, such that individuals are less likely to integrate negative feedback while learning rewarding responses. Thus, we added the learning bias parameter (LB), which

scales the learning rate α by LB when participants observe the negative feedback.

Furthermore, we captured forgetting of information across trials by also implementing decay in Q values for all cards and all choices at each trial:

$$Q^P_{t+1} = Q^P_t + \phi * (Q_0 - Q^P_t), \tag{10}$$

$$Q^L_{t+1} = Q^L_t + \phi * (Q_0 - Q^L_t) \tag{11}$$

where ϕ (the decay parameter) controls the rate of decay by pulling the Q-values towards the uniform values at which they were initialized ($Q_0 = 1/n_A$).

2.3.3 Computational model - Experiment 2

In Experiment 1, participants learned a fixed number of cards in each block (6). In Experiment 2, we varied the number of cards (or set size) between 2 and 5 in each condition, in order to manipulate involvement of working memory (A. G. E. Collins & Frank, 2012; A. G. Collins, 2017; A. G. E. Collins & Frank, 2018) and to allow us to identify WM component from RL. Hence, we extended our models by adding a working memory (WM) module (Fig. 2).

Unlike RL, WM is a one-shot learning system able to retain perfectly the previous trial's information. This is modeled by storing the immediate outcome as the stimulus-response weight.

$$W^P_{t+1}(c_t, p_t) = r_t, \tag{12}$$

$$W^L_{t+1}(c_t, l_t) = r_t \tag{13}$$

The weights stored in WM are susceptible to decay at each trial:

$$W^P_{t+1} = W^P_t + \phi * (W^{P0} - W^P_t), \tag{14}$$

$$W^L_{t+1} = W^L_t + \phi * (W^{L0} - W^L_t), \tag{15}$$

Because our Experiment 2 model combined WM and RL modules, we limited decay to WM module only. Hence, Q values used in RL computations did not decay.

We also used the softmax function to transform WM weights into choice probabilities, resembling the choice selection process in the RL module:

$$P_{wm_P} = Softmax(Beta, W^P) \tag{16}$$

(and same for label policy, using W^L).

The RL-WM model posits that choice comes from a weighted mixture of RL and WM, where one's reliance on WM is determined by the WM weight (ω) parameter:

$$P(p|c) = \omega * P_{WM}(p|c) + (1-\omega) * P_{RL}(p|c), \tag{17}$$

$$P(l|c) = \omega * P_{WM}(l|c) + (1-\omega) * P_{RL}(l|c), \tag{18}$$

where ω reflects the likelihood of an item being stored in working memory and is proportional to the ratio of capacity K and set size, scaled by the baseline propensity to rely on WM (ω_0;2):

$$\omega = min(1, \frac{K}{ns}) * \omega_0 \tag{19}$$

Importantly, the WM module could also track the other choice dimension, captured similarly as in RL with a mixture policy parameterized by ρ.

$$P_{WM}(position_i|pos.block) = \rho_P * Pwm_P(position_i) + (1-\rho_P) * Pwm_L(label(position_i)), \tag{20}$$

$$P_{WM}(label_i|lab.block) = \rho_L * Pwm_L(label_i) + (1-\rho_L) * Pwm_P(position(label_i)), \tag{21}$$

The list of parameters for RL+WM model included: learning rate (α), inverse temperature (β), lapse (ε), learning bias (LB), decay (ϕ), capacity (K), WM weight (ω),and value mixture (ρ). We explored different model variants by making different parameters fixed/varied across conditions. We explored the full model space by starting with the most complex model (all parameters varied across conditions), and gradually decreasing model complexity all the way to the simplest model (all parameters shared), while also monitoring the goodness of model fit.

2.3.4 Model fitting and comparison

Fitting Procedure. In both Experiment 1 and Experiment 2 modeling, we used maximum likelihood estimation to fit participants' individual parameters to their full sequence of choices. All parameters were bound between 0 and 1, with the exception of β parameter, which was fixed to 100 (found to improve parameter identifiability here and in previous similar tasks (Master et al., 2020)), and the K parameter of Experiment 2 models, which could take on one of the discrete values ranging from 2-5. To find the best fitting parameters, we used 20 random starting points with MATLAB's fmincon optimization function (Wilson & Collins, 2019).

Model validation. To validate whether our models could indeed capture the behavioral properties we set out to model, we simulated performance from best parameter estimates for each subject 100 times per subject. We then compared whether the model predictions from the simulated data captured the patterns we observed in the actual data set.

These simulations also allowed us to ensure that our fitting procedure could adequately recover parameters in our experimental context, by fitting the model to the simulated data and evaluating the match between the true simulation parameters and recovered parameters fit on simulated data.

Model comparison. We compared the models using Akaike Information Criterion (AIC) (Wagenmakers & Farrell, 2004) , which evaluates model fit using likelihood values and applies a complexity penalty based on the number of parameters. To ensure that our models were identifiable with AIC, we computed a confusion matrix (Wilson & Collins, 2019) by creating synthetic data sets from each model, and fitting each model to the simulated data set (computing AIC scores for each fit, see supplementary materials). This confirmed that AIC was adequately penalizing for model complexity in our situation, and more so than other criteria such as BIC.

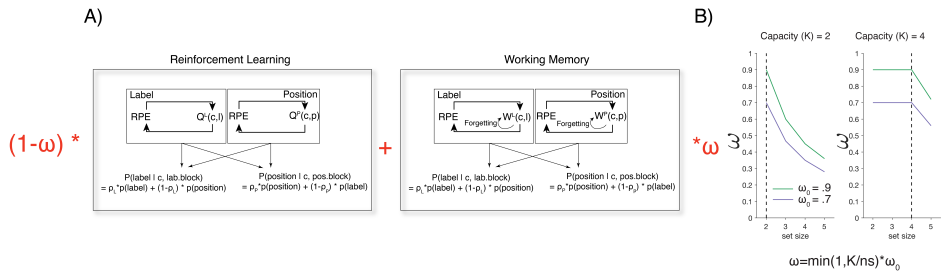


Figure 2: A) In experiment 1 we used RL model variants, which assume incremental, feedback-driven learning. In Experiment 2, we combined RL and WM modules, under the assumption that learning is a weighted interaction between RL and WM systems. B) The extent to which participants relied on WM was determined by the WM weight parameter($\boldsymbol{\omega}$), proportional to participants' WM capacity (K).

3 Results

3.0.1 Experiment 1: Behavioral results

We first asked whether participants learned differently across experimental conditions. Learning curves show that participants learned well in all conditions, as their accuracy increased with more exposure to each card (Fig. 3A). A repeated measures one-way ANOVA confirmed that there was a main effect of condition (label/position/control) on performance (F(2,61) = 97.7, p = 4.5e-26). We next tested which specific conditions contributed to the significant difference, and found that the control and position conditions were not statistically distinguishable (paired t-test: t(61)=1.61, p=.11). Thus, the additional choice feature (the labels) in the position condition did not seem to affect the choice process. The label condition performance, however, was significantly lower than the position and the control condition performances (paired t-test: position: $t(61) = 11.1$, $p = 3.8e-16$; control: $t(61) = 12.9$, $p = 5.4e-19$).

We next examined why label condition performance was worse. We hypothesized that choice was not simply noisier in the label condition, but instead that choice might be contaminated by the reward history of irrelevant motor choices. To test this hypothesis, we computed the cumulative

card-dependent label/position reward history (see methods), and quantified the proportion of error trials in which participants incorrectly chose a box with high reward history of an incorrect feature (Fig. 1B). In the position condition, participants did not make more interference errors than expected by chance level (0.5 for two possible errors) (Fig. 3B; t(61) = .13, p = .89). This confirms that the presence of labels in the position condition did not impact choice compared to the control condition. By contrast, in the label condition, the proportion of interference errors was significantly higher than chance (Fig. 3B; t(61) = 2.54, p = .01). This result suggests an asymmetry in interference between different choice spaces, in that the values of less abstract/motor action choices seem to contaminate the more abstract choice process (but not the other way around). To rule out the possibility that the effect we observe is driven by the block/condition order (i.e. transfer of incorrect strategy from the previous block), we ran a mixed-effects general linear model predicting accuracy with previous vs. current block conditions. The result of this analysis showed that participants' performance was affected by the current block condition (p = 2.22e-14), but not the previous block condition (p = 0.45), thus ruling out the order effects as possible explanation of our results.

Next, we performed a trial-by-trial analysis to examine the effect of card/label values on correct trials' reaction times (RT). For each condition, we used a mixed effects linear model to predict $\log(RT)$ from the reward history difference (RHD) between chosen and unchosen choices, where choice was label in one predictor and position in the other. The rationale behind this analysis is that if participants are engaging the appropriate decision strategy, then RTs should decrease with the higher RHD in the condition-relevant dimension (label or position), because higher RHD means greater evidence in favor of the correct response. On the other hand, in the event of interference, we expected participants' RTs to be modulated by the RHD of the incorrect dimension (e.g. position RHD in label condition). We controlled for trial number in this model, to ensure that reduction in RTs is not simply conflated with practice effect/task progression.

As predicted, participants' RTs decreased with increased respective RHD in both label (Fig. 3C; β_{label} = -.04, p = 5.1e-19), and position condition (Fig. 3C; $\beta_{position}$ = -.06, p = 3.6e-21). Label RHD did not affect the RTs in the position condition (β_{label} = -.004, p = 0.55). Hence, the mixed effects model aligned with interference errors, confirming that participants' choices were not affected by the presence of additional feature (the labels) in the po-

sition condition. On the other hand, the position RHD surprisingly increased RTs in the label condition ($\beta_{position}$ = .034, p = .001), suggesting that the interference of motor action values with label values may have resulted in the delay of choices (Fig. 3C). We compared the subject-level β estimates of the effect of incorrect dimension RHD on RTs in position and label conditions, and found that the incorrect RHD effect was significantly greater in the label condition (paired t-test: t(61) = 3.87, p = 2.6e-04), confirming the asymmetry between conditions revealed in previous analyses.

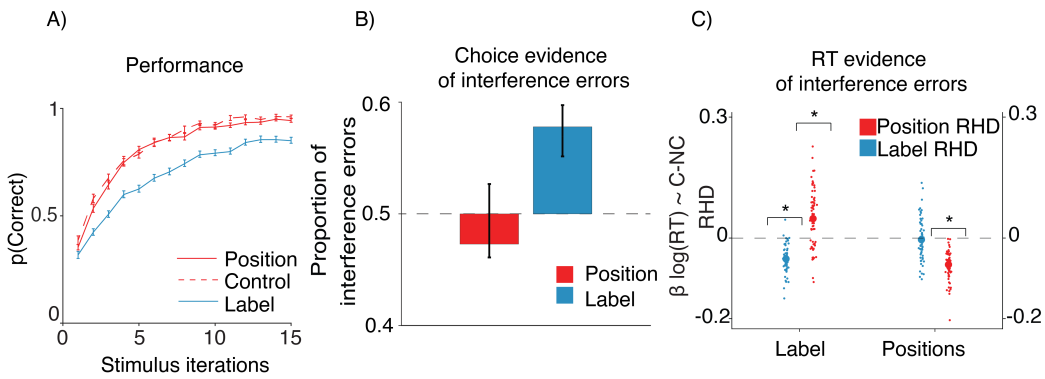


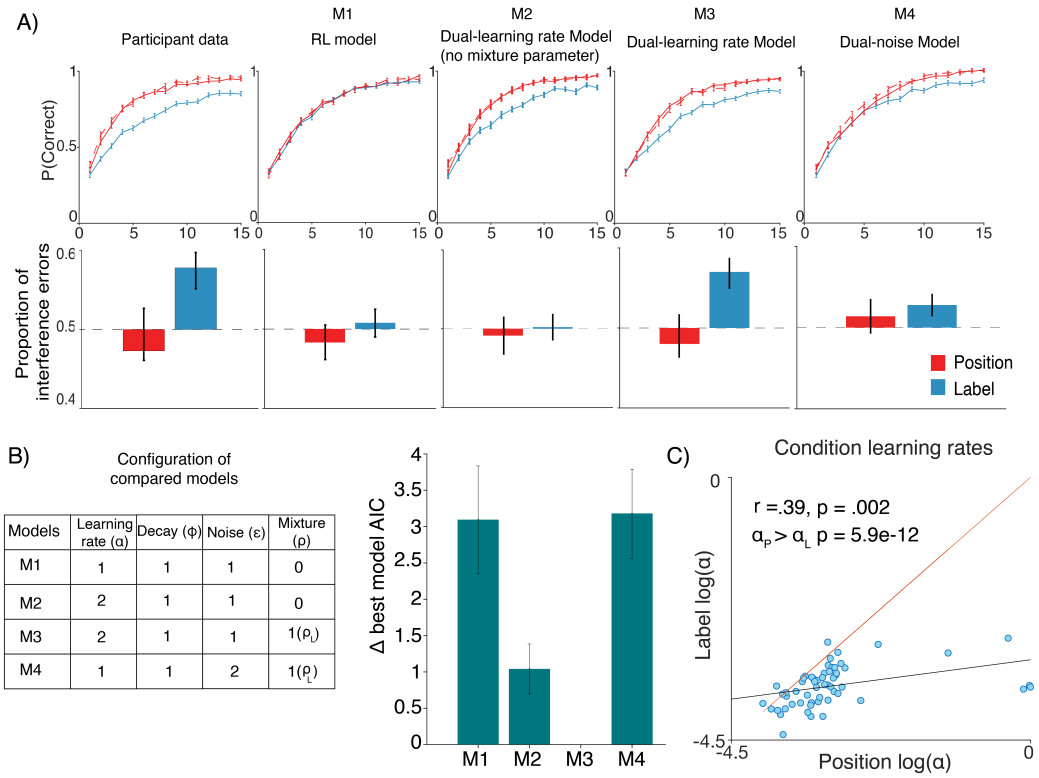
Figure 3: Experiment 1 Model-Free results. A) Participants performed worse in label condition, compared to position and control condition. Performance in position and control condition did not differ statistically. B) Asymmetric value interference: The values of motor actions interfered with values of correct labels in the label condition, thus resulting in the interference errors, but not the other way around. C) Mixed effects regression model shows that the interference of motor action reward history/values may have resulted in the delayed response/longer RTs in the label condition.

3.0.2 Experiment 1: Modeling results

We used computational modeling to tease apart the mechanisms driving the condition effects. We fit several variants of reinforcement learning (RL) models, and focus here on 4 models that represent the main different theoretical predictions (Fig. 4A; Fig. 4B). The standard RL model (M1) assumes no difference between the conditions and serves as a baseline that cannot capture the empirical effect of condition. RL model M2 lets learning rates depend on

condition, and tests the prediction that slower learning with labels is driven by different rates of reward integration. Model M3 extends model M2 with an additional mechanism, parameterized by the value mixture (ρ_L), that enables position value to influence policy in the label condition. Model M4 is an RL model with condition-dependent noise parameter (ε). M4 captures the hypothesis that the label condition is more difficult, resulting in a noisier choice process. Models M1-4 all assume $\rho_P = 1$, with no influence of labels in position blocks. Other models considered separate decay (ϕ) parameters and free position condition ρ_P, but did not improve fit (see supplementary materials).

Model M3 offered the best quantitative fit to the data, as measured by AIC (Fig. 4B). Furthermore, only model M3 was able to qualitatively reproduce patterns of behavior. Specifically, for each of the models, we simulated synthetic data sets with fit parameters and tested whether the model predictions matched the empirical results. We focused on 2 key data features in our model validation: performance averaged over the stimulus iterations (learning curves), and asymmetry in interference errors. Model validation showed that only the model with 2 learning rates and one ρ parameter (M3) captured both properties of the data (Fig. 4A). These results confirm that the learned value of (irrelevant) motor actions influenced the selection of more abstract label choices. Furthermore, model comparison results also show that slower learning in the label condition was not due to a noisier choice process, but due to reduced learning rate. Indeed, the position condition α was significantly greater than label condition α (sign test; p = 5.9e-12, Fig. 4C). Interestingly, the learning rates in the two conditions were correlated (spearman $\rho = .39, p = .003$; Fig. 4C), suggesting that the learning process in the two conditions was driven by related underlying mechanisms.



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

Figure 4: Experiment 1: Modeling Results. A) Model validation comparing the observed data to predictions of tested models. B) The list of models we compared, and the set of parameters used in each model. The AIC comparison shows that M3 fits the best. C) Comparison of condition-dependent learning rates shows that learning rates are correlated, and that label condition learning rates are significantly lower compared to position condition learning rates.

3.0.3 Experiment 2: Behavioral Results

The results of the first experiment suggest that the choice format/degree of abstraction affects learning. However, given the experimental design, our conclusions could not dissociate whether the difference in RL parameters actually reflected a difference in RL mechanisms or in WM mechanisms. Recent work (A. G. Collins, 2017; A. G. E. Collins & Frank, 2018), nevertheless, sug-

gest that RL behavior recruits other learning systems, such as WM. Hence, the variations that may appear to be driven by RL mechanisms might conceal what is actually a WM effect. To address the question of whether the choice definition matters for learning at the level of RL or WM, and whether slowed learning stems from slowed WM or RL, we ran a second experiment. In Experiment 2, we varied the number of cards (set size) to manipulate WM involvement. Furthermore, we fit a model with a WM module to test the contribution of WM mechanisms.

Experiment 2 results replicated findings from the first experiment, showing that there was a main effect of condition (Fig. 5A; repeated measures one-way ANOVA ($F(1,56) = 98.95$, $p = 5.52e\text{-}14$)). Furthermore, we replicated the pattern of interference errors suggesting that the value of less abstract choices interferes with that of more abstract choices, but not the other way around (Fig. 5B; $t(55)=2.89$; $p=.006$).

We next investigated how set size manipulation affected these results. As predicted, performance decreased with set size in both conditions (Position: $F(3,56) = 11.83$, $p = 4.51e\text{-}07$; Label: $F(3,56) = 23.498$, $p = 9.45e\text{-}13$). There was an interaction between set size and condition ($F(3,56)=16.21$, $p=5.92e\text{-}10$; Fig. 5A). The set size effect in interference errors did not reach significance ($F(3,56) = 2.17$, $p = .09$, Fig. 5C).

To better understand the source of the set size effect, we ran a general linear mixed effects model to predict trial-by-trial performance. Specifically, z-scored regressors included set size and inverse transformed delay (-1/[number of intervening trials between the current stimulus presentation, and the last rewarded presentation of the same stimulus]) to index WM contribution, and dimension-relevant, card-dependent reward history (cumulative number of earned points for each card) to assess the RL effects on performance. Lastly, we added block number as one of the regressors, in order to capture overall improvement in performance across the task. We also ran a model which tests for an interaction between individual RL/WM factors and the task condition.

A likelihood ratio test provided evidence in favor of the interaction model over a model without interactions ($p <.05$). The interaction model showed that, as expected, participants' performance increased as a function of reward history ($\beta = .62$, $p = 9.7e\text{-}30$), and decreased as a function of set size ($\beta = -.18$, $p = .00011$). There was no effect of block ($\beta = .04$, $p = .58$) or delay ($\beta = -.04$, $p = .37$), suggesting that the overall task exposure, or delay, did not affect performance over and above reward history and set size. The only

significant interaction term was the condition*reward history interaction ($\beta = .16$, $p = .01$), suggesting that the reward history contributed to an increase in performance more so in the label condition. To understand our results on a more mechanistic level, we turn to computational modeling.

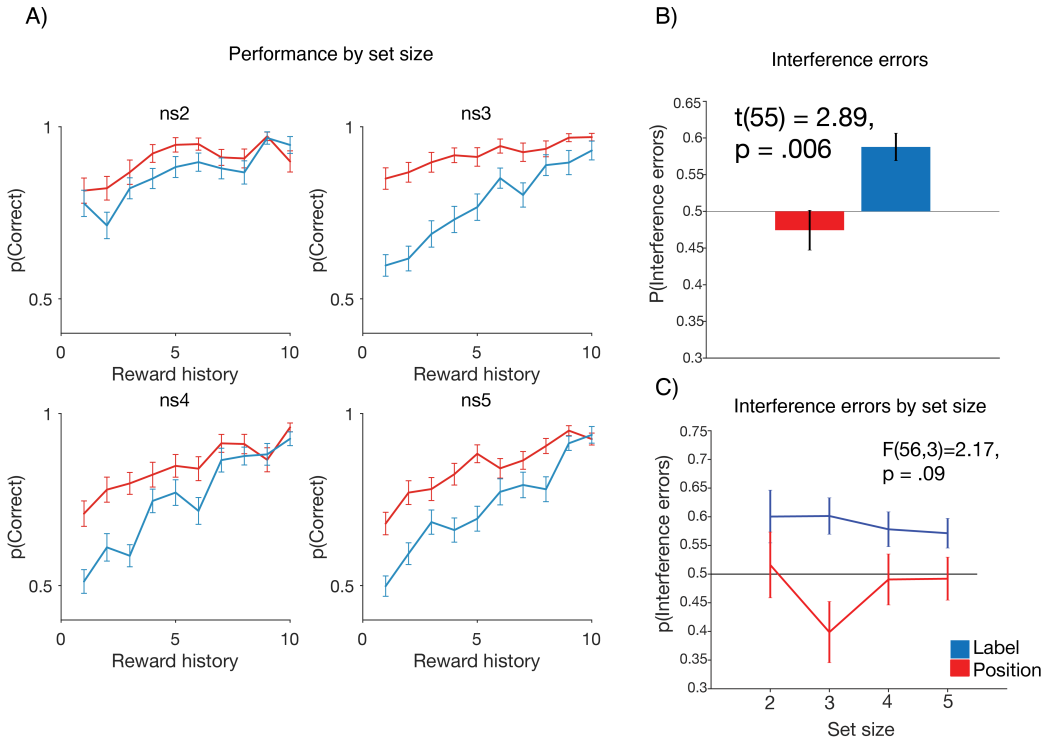


Figure 5: Experiment 2 Results. A) Participants' overall performance varied by set size (a marker of WM contribution), and was worse in the label condition. B) The asymmetry in value interference replicated from first experiment, showing that values of less abstract choice interfere with values of more abstract choices. C) The interference errors did not vary by set size.

3.0.4 Experiment 2: Modeling results

We extended the RL model by adding a separate WM module (RL+WM). This model incorporates the incremental, Q-learning aspect of RL, as well as the rapid-learning, but decay-sensitive module of WM. The final model policy is a weighted mixture of RL and WM (Fig. 2A,B), where the weighting is

proportional to one's WM capacity. In other words, the model architecture posits that if one's WM capacity is low, one might be more likely to rely on RL than WM, especially when set size (number of items) is high. We systematically varied the complexity of the RL+WM model (Fig. 2A), in order to establish whether specificity in RL or WM module parameters (or both) is necessary to capture the divergence between behavioral patterns in 2 conditions. Because the RL+WM model assumes policy for choice generation at the level of both RL and WM, we also tested if integrating irrelevant dimension interference with a ρ mixture parameter in the policy of RL module or WM module (or both) captures our data the best.

Exploring all possible models was computationally prohibitive. Thus, we explored a subset of the most relevant models. We tested models ranging from most complex (all condition-dependent parameters) to least complex (all shared parameters); (Fig. 9). Using the AIC comparison, we identified the simplest model which allowed us to capture the properties of the data (M1, (Fig. 6A)). In M1, the WM weight (ω) and ρ parameters were condition-dependent (with free ρ parameter for label condition, and position condition ρ fixed to 1). Capacity (K), learning rate (α), decay (ϕ), learning bias (LB), noise (ε) were shared across the 2 conditions - model comparison showed no benefits to making them independent (Fig. 9). We further report 3 other variants of this model: no value interference ρ (M2), ρ in RL policy alone (M3), and ρ in WM policy alone (M4) (Fig. 6A). We also report a control model with condition dependent ε and α, which would attribute the decline in label performance to noise/RL system primarily (M5). Consistent with Experiment 1 results, the AIC comparison revealed that M5 could not capture data well, and that M1 without ρ (M2) fit worse (Fig. 6A), providing additional evidence for the necessity of the interference mechanism to capture choice data, and thus, the existence of motor value interference in label blocks. However, the AIC comparison failed to significantly distinguish between the remaining models 1 (ρ in RLWM), M3 (ρ in RL) and M4 (ρ in WM) (repeated measures Anova: $F(2,56)=2.63, p=.07$), though ρ in RL models fit numerically worse, supporting the idea that we needed to include motor value interference in the WM module to account for the results. Therefore, we henceforth focus on the simplest model, M1 with condition dependent ω and ρ in RL and WM policy, as this model makes the fewest specific assumptions about RL-WM dissociation between the 2 conditions.

The M1 model adequately captured the data patterns in 1) learning curves (Fig. 6B), 2) overall interference errors (Fig. 6C) and 3) interfer-

ence errors by set size (Fig. 6D). Next, we compared condition dependent parameter. WM weight ω was significantly reduced in the label condition compared to the position condition in M1 (Fig. 6E).

Overall, the results suggested that the performance decrease in the label condition was driven primarily by deficits in WM, specifically by a decrease in the WM weight ω that indexes a priori contributions of WM to learning. Therefore, the format of the choice (more/less abstract) impacted learning, and it seemed to do so by decreasing participants' ability to use WM for learning. However, the value interference appeared to be present in both RL and WM mechanisms.

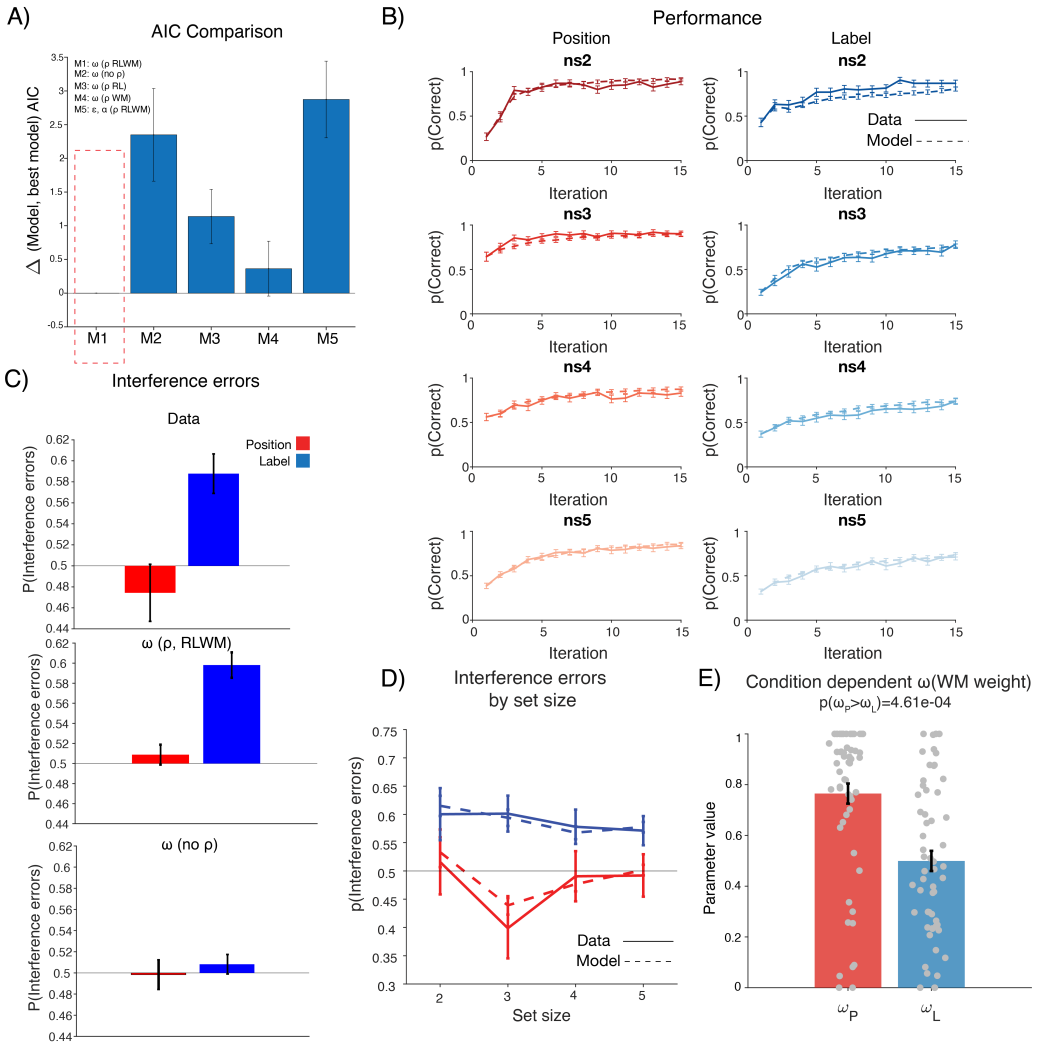


Figure 6: Experiment 2: Modeling Results. A) AIC comparison allowed us to narrow down the space of models. Models with condition-specific WM weight (ω) fit the best. Removing the mixture parameter (ρ) harmed the model fit. B) Model simulations of the best model M1 captured the behavioral data patterns. C) Model validation for M1 (ρ) and M2 (no ρ) confirms the necessity of ρ parameter in capturing the interference error patterns. D) M1 captured interference errors in different set sizes. E) Comparison of condition-dependent parameters shows that ω is lower in the label condition.

4 Discussion

Humans and animals make choices at multiple degrees of abstraction. We designed a new experimental protocol to investigate whether and how different degrees of choice abstraction impact learning. Across two experiments, behavioral analyses and computational modeling confirmed our prediction that choice abstraction impacts learning, with worse performance for more abstract choices. Computational modeling revealed two separable sources of impairment. First, value learning for more abstract relevant choices was slower, as revealed by smaller learning rates (α) in experiment 1. Second, choices were polluted by irrelevant motor values. The second experiment examined whether this dissociation originated in different neuro-cognitive systems' contributions to learning, namely RL and/or WM. Our results revealed that the reduction in learning speed for abstract choices stemmed more from WM than the RL process, with WM weight (ω) reduced but RL (α) unchanged, when controlling for WM contributions. However, the interference of low level value appeared to be present in both mechanisms. The selective reduction in WM weight implies that participants' WM resources might be leveraged to define the choice space that is then passed to both RL and WM system, but requires a higher degree of computation when it comes to a more abstract choice space, thus leaving reduced resources for actual learning.

In both experiments, we found an asymmetry in interference between choice types. When participants learned to make more abstract choices (selecting a label) that required a subsequent motor action (pressing the key corresponding to the label's location), their choices were influenced by the (irrelevant) reward history of motor actions. By contrast, when participants learned to make less abstract choices (pressing the key corresponding to the box location), they were not influenced by the irrelevant reward history of label boxes. This result is consistent with a choice hierarchy interpretation, where participants may be unable to turn off credit assignment to irrelevant choice dimensions when the realization of their (abstract) choice does involve this dimension. By contrast, credit assignment can be turned off for more abstract choice dimensions that are not task-relevant.

While our evidence implies that participants exhibit a decision bias towards motor actions, we acknowledge that our protocol cannot disambiguate between the motor actions themselves and the corresponding spatial location of the boxes. That is, we cannot confirm whether the participants track the

value of specific motor actions (index/middle/ring finger key press) or of the corresponding box positions (left/middle/right). Hence, a competing interpretation of our results would be that spatial position, rather than motor actions, are prioritized in tracking value, compared to other visual features such as labels. To completely rule out this possibility, we would need to modify the current task with a condition where the motor actions are not aligned with the specific positions, and inspect whether the interference effect persists in such condition. However, we think this account is less likely than a choice abstraction account, which explains our results more parsimoniously, without requiring a "special status" for a "position" visual feature.

Furthermore, animal research supports this interpretation, as it shows differences in the neural code of choices, which are defined primarily as motor actions versus more abstract choices (Rothenhoefer et al., 2017; Luk & Wallis, 2013). Specifically, these studies have utilized recordings from neurons of animals trained to perform a task that contrasted motor-action choices with stimulus goal choices, in order to identify the neural substrates which differentiate between the two. The results seem to implicate prefrontal cortex (PFC), anterior cingulate cortex (ACC) orbitofrontal cortex (OFC) and striatal regions (ventral striatum) as areas that differentiate between how choices with different levels of abstraction are coded in the brain. Therefore, it is likely that it truly is a dissociation between motor actions, rather than positions, and more abstract choices that led to the interference and the effects we observed in our work.

In addition, frontal areas, primarily PFC, and basal ganglia (BG), are also frequently investigated as neural mechanisms which support hierarchical reasoning/learning (A. G. E. Collins & Frank, 2013). Previous research offers converging insights, suggesting that the cortico-BG loops support representations of both low-level associations and abstract rules/task sets, giving rise to latent representations that can be used to accelerate learning in novel sittings (A. G. E. Collins & Frank, 2013; Stolle & Precup, 2002; Xia & Collins, 2020; Eckstein et al., 2019).

Both experiments implicated overall slowed learning, in addition to value interference, in the worse performance for more abstract choices. Our first experiment (which allowed us to test RL models only) implicated the learning rate (usually interpreted as a marker of RL system (Eckstein et al., 2021)) as the mechanism driving the difference between conditions with different levels of choice abstraction. However, our second experiment enabled us to test the more holistic hybrid model of RL and WM, and revealed that the im-

pairment in the more abstract choice condition likely stemmed from the WM system, rather than RL. Previous work has shown that executive function (EF), in its different forms, contributes to RL computations (A. G. Collins, 2017; Niv, 2019). The general summary of this work is that high-dimensional environments/tasks pose difficulty to RL; EF then acts as an information compressor, making the information processing more efficient for RL (Rmus et al., 2020). More abstract choice conditions might more heavily rely the contribution of EF (in this case WM) relative to the less abstract condition. Therefore, resource-limited WM might be leveraged to define the choice space (i.e. relevant features of the choice space-labels in label condition). As a result, the WM weight included in the WM + RL hybrid model, which indexes the WM contribution to learning, appear reduced in the label condition. Our interpretation of this result is that this reduction in WM contribution might indicate that some of participants' limited WM resources are recruited elsewhere, and specifically that it has already been used to define the choice space over which learning and decision making occurs.

While we conclude that WM is used for defining the choice space, consistent with prior results on EF contributions to RL computations (Todd, Niv, & Cohen, n.d.), we do not make any particular assumptions about how the use of choice space is divided between RL and WM once it's defined. We tested different model variations, with the parameter that mixes label/position values, to explain value interference at the policy level of RL, WM or both. If there was clear evidence in favor of the mixture parameter in RL or WM policy, it would imply that the policy generation based on choice space is primarily driven by that system. However, our model comparison revealed no evidence that the mixture parameter is specific to either RL or WM, suggesting that the choice space is shared between the two. This will be important to confirm in future research.

A competing interpretation for our findings of slowed learning for more abstract choices is that the label condition required more attention and was more difficult. While this is true, we took steps to mitigate this potential confound on two levels - task design and modeling. In the task design, we constructed the single trial structure such that participants had a chance to see box labels first, before the onset of the card. By doing this we aimed to eliminate potential advantages of the position condition, where participants do not need to perform an additional process of identifying the label location prior to executing the response. Furthermore, our modeling enabled us to validate the effects of our task design. Specifically, in both experiments we

tested the model with condition-dependent noise parameters, which predicts that different noise/difficulty levels are what drive the performance difference in our conditions. This model did not fit the data well (Experiment 1: Best model AIC > 2 noise model AIC t(56)= -5.179, p= 3.13e-06; Experiment 2: Best model AIC > 2 noise model AIC t(56)= -5.05, p= 4.98e-06), making it unlikely that difficulty-induced lack of attention/motivation could explain our condition effect.

Surprisingly, we found that participants' response times (RT) on correct trials increased as a function of position reward history difference (RHD) in the label condition. This implies that when both label and position sorting rules were in agreement as to the best choice to make (i.e. the blue box was the correct box, and was in the position that had been most rewarded so far), response times tend to be longer (the corresponding effect was not observed in the position condition, where label RHD had no effect on RTs). This is, therefore, a counterintuitive effect, as we would expect the congruent information to accelerate response execution, rather than slow it, as observed here. One possibility might be that participants do engage in a form of arbitration between selection of different response types. Specifically, they might be biased to execute the motor action based on the reward history difference, as it seems to present itself as a default option based on our results. However, because they are informed that the more abstract response selection is correct for the given block, they might delay the response execution, in order to override the default. Nevertheless, this is a speculation - careful modeling of response times is required to further explain this effect, which is beyond the scope of this paper. This account would also predict the highest degree of conflict in this congruent situation, rather than in situations where both rules disagree. It will be an important question to solve in future research.

Our results highlight the importance of correct credit assignment, and investigation of mechanisms which might lead to errors in the credit assignment process. Our results are consistent with the previous research suggesting that motor actions might have a stronger effect on the choice selection process than is usually considered (Shahar et al., 2019). Our modeling approach allowed us to show that the mixture of Q values at the policy level is what might lead to the interference effect/incorrect credit assignment. However, as of now, we cannot conclusively say whether the mixture happens selectively at the policy level of RL, WM or both.

Identification of correct rewarding responses is a critical building block

of adaptive/goal-directed behavior. Impairments in one's ability to identify the appropriate choice space, which is then used for inference process, may consequently result in maladaptive/suboptimal behavioral patterns. Our interference effect results suggest that some aspects of the choice space might incorrectly be overvalued, thus resulting in choice patterns that reflect repeated erroneous selection of incorrect choice types, or an inability to utilize flexible stimulus-response mapping. These kinds of perseverative responses are reminiscent of the inability to disengage from certain actions, observed in conditions such as obsessive-compulsive disorder (OCD) (Rosa-Alcázar et al., 2020). It would be interesting to use our task and computational modeling approach to investigate whether the mixture/interference of values at the policy level could also explain the behavior of such population.

In conclusion, our findings provide evidence that the degree of choice abstraction and how we define a choice have important implications for the learning process. The behavioral patterns (i.e. value interference from less abstract choices) are consistent with the premises of hierarchy in learning and behavior (i.e. lower levels in hierarchy impacting processing in higher levels), which has become an increasingly promising topic of research (Eckstein & Collins, 2019; A. G. E. Collins & Frank, 2013; Stolle & Precup, 2002). We also demonstrate additional evidence, relevant to the definition of the choice space, that EF (specifically WM) contributes to RL in reward-driven behaviors (Rmus et al., 2020), further demonstrating the complex interplay between various neuro-cognitive systems.

5 Acknowledgments

This work was funded by NSF2020844 to AGEC.

6 References

References

Ballard, I., Miller, E. M., Piantadosi, S. T., Goodman, N. D., & McClure, S. M. (2018, November). Beyond Reward Prediction Errors: Human Striatum Updates Rule Values During Learning. Cerebral Cor-

tex, 28(11), 3965–3975. Retrieved 2021-07-28, from https://doi.org/10.1093/cercor/bhx259 doi: 10.1093/cercor/bhx259

Bornstein, A. M., & Daw, N. D. (2013, December). Cortical and Hippocampal Correlates of Deliberation during Model-Based Decisions for Rewards in Humans. PLOS Computational Biology, 9(12), e1003387. Retrieved 2021-07-11, from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003387 (Publisher: Public Library of Science) doi: 10.1371/journal.pcbi.1003387

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017, December). Reminders of past choices bias decisions for reward in humans. Nature Communications, 8(1), 15958. Retrieved 2021-07-22, from http://www.nature.com/articles/ncomms15958 doi: 10.1038/ncomms15958

Collins, A. G. (2017, December). The tortoise and the hare: interactions between reinforcement learning and working memory (preprint). Neuroscience. Retrieved 2021-07-07, from http://biorxiv.org/lookup/doi/10.1101/234724 doi: 10.1101/234724

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017, April). Working Memory Load Strengthens Reward Prediction Errors. Journal of Neuroscience, 37(16), 4332–4342. Retrieved 2021-02-08, from https://www.jneurosci.org/content/37/16/4332 (Publisher: Society for Neuroscience Section: Research Articles) doi: 10.1523/JNEUROSCI.2700-16.2017

Collins, A. G. E., & Frank, M. J. (2012, April). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis: Working memory in reinforcement learning. European Journal of Neuroscience, 35(7), 1024–1035. Retrieved 2021-07-07, from http://doi.wiley.com/10.1111/j.1460-9568.2011.07980.x doi: 10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2013, January). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. Psychological Review, 120(1), 190–229. Retrieved 2021-02-08, from http://doi.apa.org/getdoi.cfm?doi=10.1037/a0030852 doi: 10.1037/a0030852

Collins, A. G. E., & Frank, M. J. (2018, March). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proceedings of the National Academy of Sciences, 115(10), 2502–2507. Retrieved 2021-02-08,

from https://www.pnas.org/content/115/10/2502 (Publisher: National Academy of Sciences Section: Biological Sciences) doi: 10.1073/pnas.1720963115

Daw, N., Gershman, S., Seymour, B., Dayan, P., & Dolan, R. (2011, March). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron, 69(6), 1204–1215. Retrieved 2021-02-08, from https://www.sciencedirect.com/science/article/pii/S0896627311001255 doi: 10.1016/j.neuron.2011.02.027

de Leeuw, J. R. (2015, March). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behavior Research Methods, 47(1), 1–12. Retrieved 2021-08-25, from http://link.springer.com/10.3758/s13428-014-0458-y doi: 10.3758/s13428-014-0458-y

Eckstein, M. K., & Collins, A. G. (2019, August). Computational evidence for hierarchically-structured reinforcement learning in humans. bioRxiv, 731752. Retrieved 2021-02-08, from https://www.biorxiv.org/content/10.1101/731752v1 (Publisher: Cold Spring Harbor Laboratory Section: New Results) doi: 10.1101/731752

Eckstein, M. K., Starr, A., & Bunge, S. A. (2019, April). How the inference of hierarchical rules unfolds over time. Cognition, 185, 151–162. Retrieved 2021-02-08, from https://www.sciencedirect.com/science/article/pii/S0010027719300150 doi: 10.1016/j.cognition.2019.01.009

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021, October). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. Current Opinion in Behavioral Sciences, 41, 128–137. Retrieved 2021-07-26, from https://linkinghub.elsevier.com/retrieve/pii/S2352154621001236 doi: 10.1016/j.cobeha.2021.06.004

Farashahi, S., Rowe, K., Aslami, Z., Lee, D., & Soltani, A. (2017, November). Feature-based learning improves adaptability without compromising precision. Nature Communications, 8(1), 1768. Retrieved 2021-02-08, from https://www.nature.com/articles/s41467-017-01874-w (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41467-017-01874-w

Foerde, K., & Shohamy, D. (2011, September). Feedback Timing Modulates Brain Systems for Learning in Humans. Journal of Neuroscience, 31(37), 13157–13167. Retrieved 2021-07-04, from https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2701-11.2011 doi: 10.1523/JNEUROSCI.2701-11.2011

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007, October). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proceedings of the National Academy of Sciences, 104(41), 16311–16316.

Gershman, S. J. (2015, October). Do learning rates adapt to the distribution of rewards? Psychonomic Bulletin & Review, 22(5), 1320–1327. Retrieved 2021-07-07, from http://link.springer.com/10.3758/s13423-014-0790-3 doi: 10.3758/s13423-014-0790-3

Luk, C.-H., & Wallis, J. D. (2013, January). 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. E. (2020, February). Disentangling the systems contributing to changes in learning during adolescence. Developmental Cognitive Neuroscience, 41, 100732. Retrieved 2021-02-08, from https://www.sciencedirect.com/science/article/pii/S1878929319303196 doi: 10.1016/j.dcn.2019.100732

McDougle, S. D., Boggess, M. J., Crossley, M. J., Parvin, D., Ivry, R. B., & Taylor, J. A. (2016, June). Credit assignment in movement-dependent reinforcement learning. Proceedings of the National Academy of Sciences, 113(24), 6797–6802. Retrieved 2021-06-15, from https://www.pnas.org/content/113/24/6797 (Publisher: National Academy of Sciences Section: Biological Sciences) doi: 10.1073/pnas.1523669113

Nassar, M. R., & Frank, M. J. (2016, October). Taming the beast: extracting generalizable knowledge from computational models of cognition. Current opinion in behavioral sciences, 11, 49–54. Retrieved 2021-07-07, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001502/ doi: 10.1016/j.cobeha.2016.04.003

Niv, Y. (2019, October). Learning task-state representations. Nature Neuroscience, 22(10), 1544–1553. Retrieved 2021-02-08, from https://www.nature.com/articles/s41593-019-0470-8 (Number: 10 Publisher: Nature Publishing Group) doi: 10.1038/s41593-019-0470-8

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012, January). Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. Journal of Neuroscience, 32(2), 551–562.

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., & Gluck, M. A. (2001, November). Interactive memory systems in the human brain. Nature, 414(6863), 546–550. Re-

trieved 2021-07-28, from https://www.nature.com/articles/35107080 (Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6863 Primary_atype: Research Publisher: Nature Publishing Group) doi: 10.1038/35107080

Rescorla, R. A., & Solomon, R. L. (1967, May). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. Psychological Review, 74(3), 151–182. doi: 10.1037/h0024475

Rmus, M., McDougle, S., & Collins, A. (2020, June). The Role of Executive Function in Shaping Reinforcement Learning (Tech. Rep.). PsyArXiv. Retrieved 2021-07-10, from https://psyarxiv.com/9cvw3/ (type: article) doi: 10.31234/osf.io/9cvw3

Rosa-Alcázar, , Olivares-Olivares, P. J., Martínez-Esparza, I. C., Parada-Navas, J. L., Rosa-Alcázar, A. I., & Olivares-Rodríguez, J. (2020). Cognitive flexibility and response inhibition in patients with Obsessive-Compulsive Disorder and Generalized Anxiety Disorder. International Journal of Clinical and Health Psychology : IJCHP, 20(1), 20–28. Retrieved 2021-07-26, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6994753/ doi: 10.1016/j.ijchp.2019.07.006

Rothenhoefer, K. M., Costa, V. D., Bartolo, R., Vicario-Feliciano, R., Murray, E. A., & Averbeck, B. B. (2017, July). Effects of Ventral Striatum Lesions on Stimulus-Based versus Action-Based Reinforcement Learning. The Journal of Neuroscience, 37(29), 6902–6914. Retrieved 2021-07-07, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5518420/ doi: 10.1523/JNEUROSCI.0631-17.2017

Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., & Dolan, R. J. (2019, August). Credit assignment to state-independent task representations and its relationship with model-based decision making. Proceedings of the National Academy of Sciences, 116(32), 15871–15876.

Stolle, M., & Precup, D. (2002). Learning Options in Reinforcement Learning. In Lecture Notes in Computer Science (pp. 212–223).

Sugawara, M., & Katahira, K. (2021, February). Dissociation between asymmetric value updating and perseverance in human reinforcement learning. Scientific Reports, 11(1), 3574. Retrieved 2021-07-07, from https://www.nature.com/articles/s41598-020-80593-7 (Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cognitive neuroscience;Deci-

sion Subject_term_id: cognitive-neuroscience;decision) doi: 10.1038/s41598-020-80593-7

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning, second edition: An Introduction. MIT Press. (Google-Books-ID: uWV0DwAAQBAJ)

Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A., & Wilbrecht, L. (2012, September). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nature Neuroscience, 15(9), 1281–1289. Retrieved 2021-07-18, from https://www.nature.com/articles/nn.3188 (Bandiera_abtest: a Cg_type: Nature Research Journals Number: 9 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Decision;Neuronal physiology Subject_term_id: decision;neuronal-physiology) doi: 10.1038/nn.3188

Todd, M. T., Niv, Y., & Cohen, J. D. (n.d.). Learning to use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. , 8.

Vikbladh, O. M., Meager, M. R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., … Daw, N. D. (2019, May). Hippocampal Contributions to Model-Based Planning and Spatial Memory. Neuron, 102(3), 683–693.e4. Retrieved 2021-07-22, from https://linkinghub.elsevier.com/retrieve/pii/S0896627319301230 doi: 10.1016/j.neuron.2019.02.014

Wagenmakers, E.-J., & Farrell, S. (2004, February). AIC model selection using Akaike weights. Psychonomic Bulletin & Review, 11(1), 192–196. Retrieved 2021-07-07, from http://link.springer.com/10.3758/BF03206482 doi: 10.3758/BF03206482

Wilson, R. C., & Collins, A. G. (2019, November). Ten simple rules for the computational modeling of behavioral data. eLife, 8, e49547. Retrieved 2021-06-28, from https://doi.org/10.7554/eLife.49547 (Publisher: eLife Sciences Publications, Ltd) doi: 10.7554/eLife.49547

Wimmer, G. E., & Shohamy, D. (2012, October). Preference by Association: How Memory Mechanisms in the Hippocampus Bias Decisions. Science, 338(6104), 270–273. Retrieved 2021-07-11, from https://science.sciencemag.org/content/338/6104/270 (Publisher: American Association for the Advancement of Science Section: Report) doi: 10.1126/science.1223252

Xia, L., & Collins, A. G. E. (2020, February). Temporal and state abstractions for efficient learning, transfer and composition in hu-

mans. bioRxiv, 2020.02.20.958587. Retrieved 2021-02-08, from https://www.biorxiv.org/content/10.1101/2020.02.20.958587v1 (Publisher: Cold Spring Harbor Laboratory Section: New Results) doi: 10.1101/2020.02.20.958587

7 Supplementary materials

Experiment 1 additional model comparisons. We tested whether an additional decay parameter or an additional mixture parameter improved the fit to the data. These models did not improve the fit compared to M3 (our winning model).

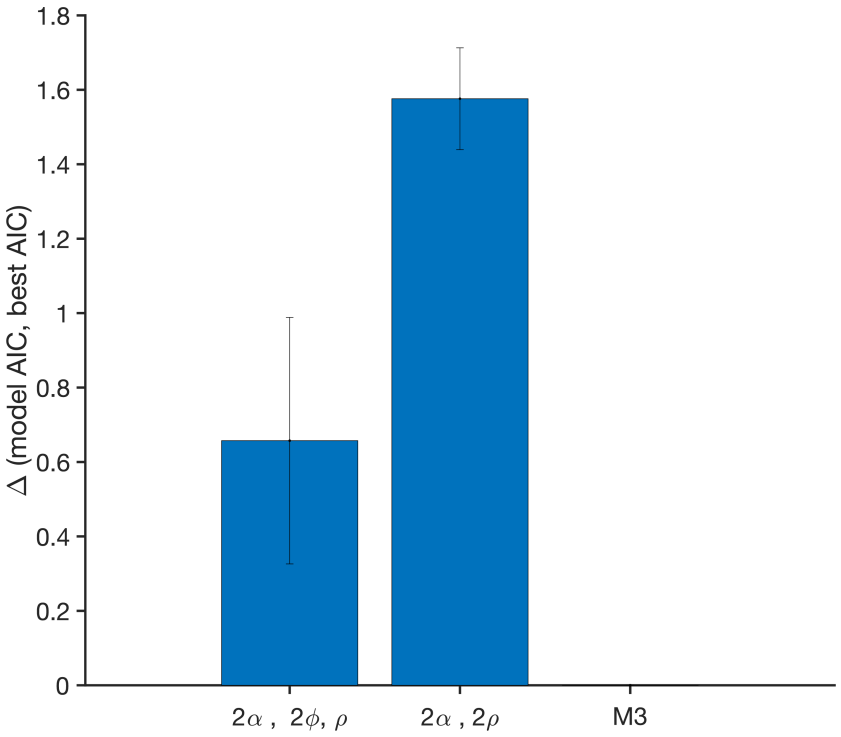


Figure 7: Additional models tested in Experiment 1.

Experiment 1 confusion matrix. To demonstrate the identifiability of our models (i.e. models are meaningfully different from one another), we sim-

ulated the data from each model on 62 iterations (number of participants). We used best parameter estimates for each participant to create a synthetic data set on each iteration. We then fitted each of the models to each simulated data set with 20 random starting points, to match the fitting procedure to participants' data. Next, we computed the proportion of the times each model fit the best. If the models are identifiable, the model the data was simulated from should fit the best on most iterations (i.e. the matrix should have the highest proportion of best fit values on its diagonal). The confusion matrix showed that our models are identifiable.

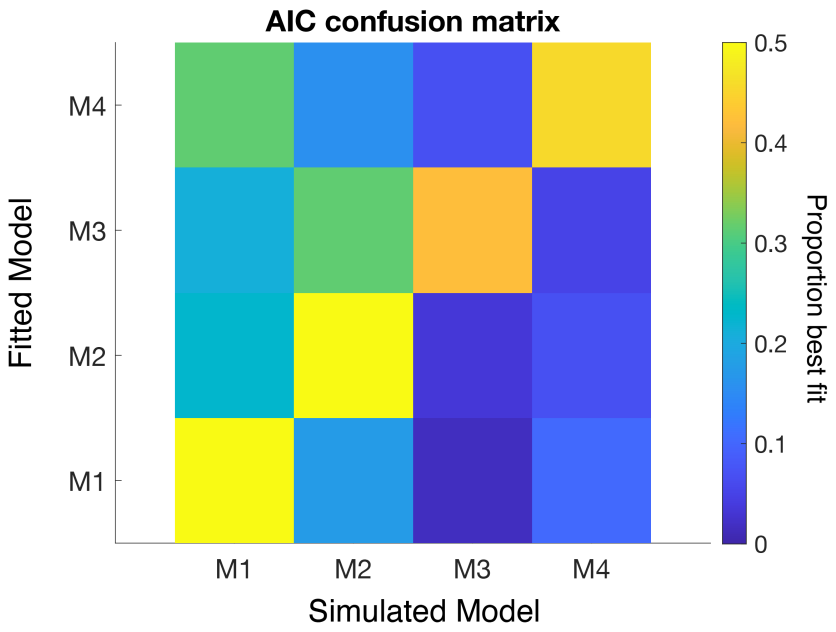


Figure 8: Confusion matrix of the main models tested in Experiment 1.

In our second experiment, we fit a considerable range of models, starting with the most complex (all RL + WM parameters condition-dependent), to the simplest (all parameters shared across conditions). We systematically varied the complexity of the model, while monitoring the model fit/complexity tradeoff using AIC scores, in order to test which parameters are necessary for capturing the difference between the conditions while also making sure our models are not overfitting (Fig. 9).

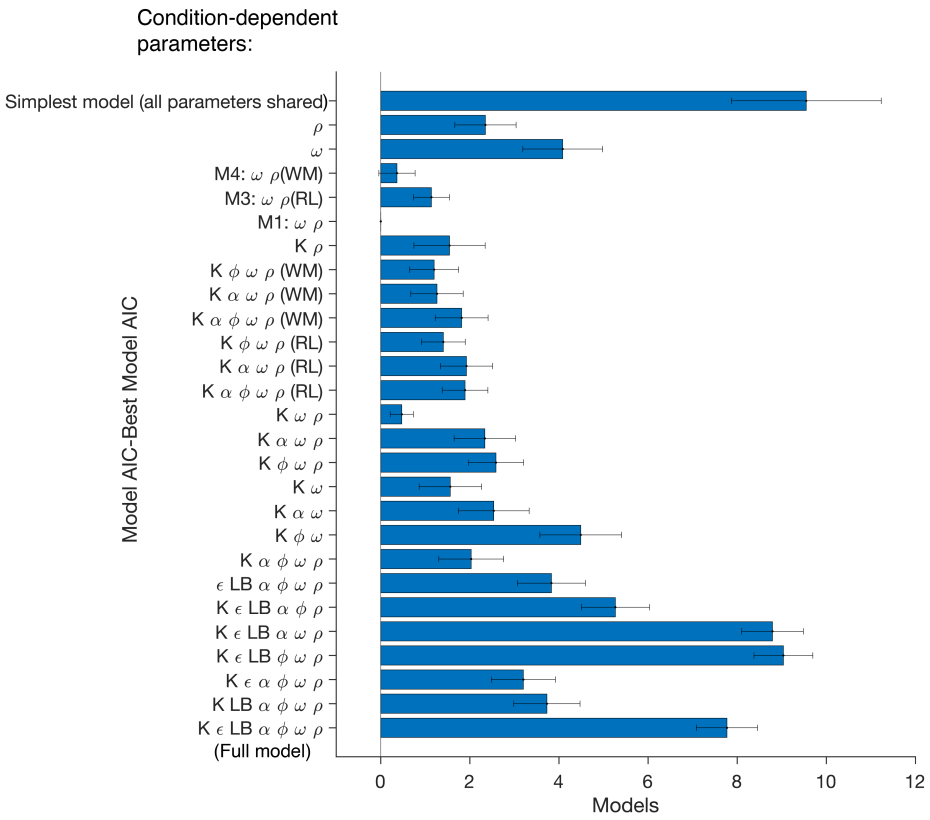


Figure 9: AIC comparison of models tested in Experiment 2. Here we show the difference in individual AIC scores between M3, and all other models that were tested.

Experiment 2 Confusion Matrix. We tested the identifiability of our models in experiment 2 by creating a confusion matrix (see Experiment 1 confusion matrix section in supplementary material for details on the procedure). We constructed two different confusion matrices, which test for identifiability of our model along 2 different dimensions. Our first confusion matrix allowed us to test whether the models with different placements of the ρ parameter (i.e. with wrong choice dimension policy mixture in RL, WM or both) are meaningfully dissociable. The confusion matrix shows that the models with mixture ρ in RL and WM policy can be dissociated (Fig. 10). The data simulated from the model with ρ parameter in both WM and RL policy was

fit equally well by that model and the model with ρ in WM policy alone. This is consistent with our results, as model comparison revealed that AIC scores did not meaningfully differ between these two models.

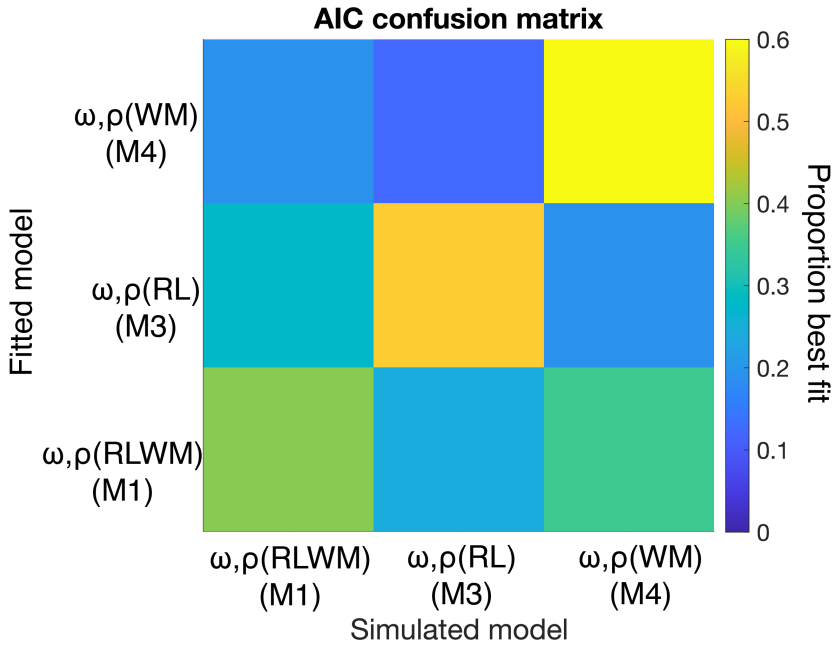


Figure 10: Confusion Matrix 1. Proportion of times the models fitted different simulated data sets best, based on cross-fit AIC scores for models with different placement of ρ paramater.

Our second confusion matrix tested whether we can dissociate the model we converged on in the main text (M1, ω with RL+WM ρ) from variations of model with 1) no ρ parameter, and 2) shared WM weight ω. Our results showed that our models are mostly identifiable, with an exception of M2 (Fig. 11). However, M2 cannot produce the observed qualitative error patterns, providing another method to rule out this model.

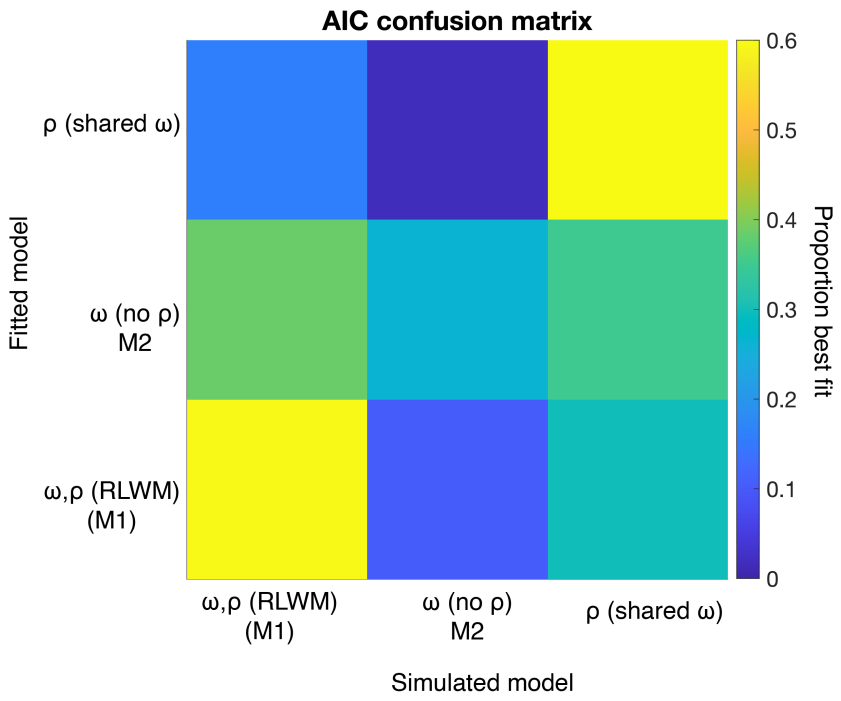


Figure 11: Confusion Matrix 2. Proportion of times the models fitted different simulated data sets best, based on cross-fit AIC scores for models with condition dependent **ρ** and **ω** parameters (M1), condition dependent **ω** (M2), and condition dependent **ρ**.