Learning at Variable Attentional Load Requires Cooperation of Working Memory, Meta-learning, and Attention-augmented Reinforcement Learning

Thilo Womelsdorf[1], **Marcus R. Watson**[2], **and Paul Tiesinga**[3]

Abstract

■ Flexible learning of changing reward contingencies can be realized with different strategies. A fast learning strategy involves using working memory of recently rewarded objects to guide choices. A slower learning strategy uses prediction errors to gradually update value expectations to improve choices. How the fast and slow strategies work together in scenarios with real-world stimulus complexity is not well known. Here, we aim to disentangle their relative contributions in rhesus monkeys while they learned the relevance of object features at variable attentional load. We found that learning behavior across six monkeys is consistently best predicted with a model combining (i) fast working memory and (ii) slower reinforcement learning from differently weighted positive and negative prediction errors as well as (iii) selective suppression of nonchosen feature values and (iv) a meta-learning mechanism that enhances exploration rates based on a memory trace of recent errors. The optimal model parameter settings suggest that these mechanisms cooperate differently at low and high attentional loads. Whereas working memory was essential for efficient learning at lower attentional loads, enhanced weighting of negative prediction errors and meta-learning were essential for efficient learning at higher attentional loads. Together, these findings pinpoint a canonical set of learning mechanisms and suggest how they may cooperate when subjects flexibly adjust to environments with variable real-world attentional demands. ■

INTRODUCTION

Cognitive flexibility is realized through multiple mechanisms (Dajani & Uddin, 2015), including recognizing that environmental demands change, the rapid updating of expectations, and the shifting of response strategies away from irrelevant toward newly relevant information. The combination of these processes is a computational challenge as they operate on different time scales ranging from slow integration of reward histories to faster updating of expected values given immediate reward experiences (Botvinick et al., 2019). How fast and slow learning processes cooperate to bring about efficient learning is not well understood.

Fast adaptation to changing reward contingencies depends on a fast learning mechanism. Previous studies suggest that such a fast learning strategy can be based on different strategies. One strategy involves memorizing successful experiences in a working memory (WM) and guiding future choices to those objects that have highest expected reward value in WM (Alexander & Womelsdorf, 2021; McDougle & Collins, 2020; Viejo, Girard, Procyk, & Khamassi, 2018; Alexander & Brown, 2015; Collins, Brown, Gold, Waltz, & Frank, 2014; Collins & Frank, 2012). This WM strategy is similar to recent "episodic" learning models that store instances of episodes as a means to increase learning speed when similar episodes are encountered (Botvinick et al., 2019; Gershman & Daw, 2017).

A second fast learning mechanism uses an attentional strategy that enhances learning from those experiences that were selectively attended (Oemisch et al., 2019; Niv et al., 2015; Rombouts, Bohte, & Roelfsema, 2015). The advantage of this strategy is an efficient sampling of values when there are many alternatives or uncertain reward feedback (Farashahi, Rowe, Aslami, Lee, & Soltani, 2017; Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017; Kruschke, 2011). Empirically, such an attentional mechanism accounts for learning values of objects and features within complex multidimensional stimulus spaces (Hassani et al., 2017; Leong et al., 2017; Niv et al., 2015; Wilson & Niv, 2011). In these multidimensional spaces, learning from sampling all possible object instances can be impractical and slows down learning to a greater extent than what is observed in humans and monkeys (Oemisch et al., 2019; Farashahi, Rowe, et al., 2017). Instead, learners appear to speed up learning by learning stronger from objects that are attended and actively chosen, while penalizing features associated with nonchosen objects (Oemisch et al., 2019; Hassani et al., 2017; Leong et al., 2017; Niv et al., 2015; Wilson & Niv, 2011).

In addition to WM and attention-based strategies, various findings indicate that learning can be critically enhanced by selectively increasing the rate of exploration during difficult or volatile learning stages (Soltani & Izquierdo,

[1] Vanderbilt University, [2] York University, Toronto, ON, Canada, [3] Radboud University Nijmegen

https://doi.org/10.1162/jocn_a_01780

2019; Khamassi, Quilodran, Enel, Dominey, & Procyk, 2015). Such a meta-learning strategy, for example, increases the rate of exploring options as opposed to exploiting previously learned value estimates (Tomov, Truong, Hundia, & Gershman, 2020). This and other meta-learning approaches have been successfully used to account for learning rewarded object locations in monkeys (Khamassi et al., 2015) and for speeding up learning of multiarm bandit problems (Wang et al., 2018).

There is evidence for all three proposed strategies in learning, but only few empirical studies characterize the contribution of different learning strategies. Thus, it is unknown whether WM, attention-augmented reinforcement learning (RL), and meta-learning approaches are all used during learning in differently complex environments and whether they differently cooperate at low and high learning difficulty.

To address this issue, we set out to test and disentangle the specific contribution of various computational mechanisms for flexibly learning the relevance of visual object features. We trained six monkeys to learn the reward value of object features in environments with varying numbers of irrelevant distracting feature dimensions. By increasing the number of distracting features, we increased attentional load, which resulted in successively slower learning behavior. We found that, across monkeys, learning speed was best predicted by a computational RL model that combines WM, attention-augmented RL, a separate learning rate for erroneous choices, and meta-learning. The optimal model parameter settings, which account for a significant fraction of the observed choices, suggest that the contribution of these individual learning mechanisms varied systematically with attentional load. WM contributed to learning speed particularly at low and medium loads, meta-learning contributed maximal at high loads, whereas selective decay of nonattended feature values was an essential learning mechanism across all attentional loads.

METHODS

Experimental Design

Six male macaque monkeys, age ranging from 6 to 9 years and weighing 8.5–14.4 kg, performed the experiments. All animal and experimental procedures were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and the Society for Neuroscience Guidelines and Policies and approved by the Vanderbilt University Institutional Animal Care and Use Committee.

The experiment was controlled by the Unified Suite for Experiments using the Unity 3-D game engine for behavioral control and visual display (Watson, Voloh, Thomas, Hasan, & Womelsdorf, 2019). Four animals performed the experiment in cage-based touchscreen Kiosk Testing Stations described in Womelsdorf et al. (2021), and two animals performed the experiment in a sound-attenuating experimental booth. All experiments used 3-D rendered objects, so-called Quaddles (Watson, Voloh, Naghizadeh, & Womelsdorf, 2019), that were defined by their body shape, arm style, surface pattern, and color (Figure 1A). We used up to nine possible body shapes, six possible colors, 11 possible arm types, and nine possible surface patterns as feature values. The six colors were equidistant within the perceptually defined color space CIELAB. Objects extended ~3 cm on the screen corresponding to ~2.5° of visual angle and were presented either on a 24-in. BenQ monitor or an Elo 2094L 19.5 LCD touchscreen running at a 60-Hz refresh rate with a 1920 × 1080 pixel resolution.

Task Paradigm

Animals performed a feature–reward learning task that required learning through trial-and-error which feature of multidimensional objects is associated with reward. The feature that was rewarded, that is, the feature–reward rule, stayed constant for blocks of 35–60 trials and then switched randomly to another feature (Figure 1B). Individual trials (Figure 1C) were initiated by either touching a central blue square (four monkeys) or fixating the blue square for 0.5 sec. After a 0.3-sec delay, three objects were presented at the corners of a virtual square grid spanning 15 cm on the screen (~24°). The animals had up to 5 sec to choose one object by touching it for 0.1 sec (four monkeys) or maintaining gaze at an object for 0.7 sec (two monkeys). After the choice of an object, visual feedback was provided as a colored disk behind the selected object (yellow/gray for rewarded/not rewarded choices, respectively) concomitant with auditory feedback (low/high-pitched sound for nonrewarded/rewarded choices, respectively). Choices of the object with the rewarded feature resulted in fluid reward of 0.3 sec after the onset of the visual and auditory feedback.

For each learning block, a unique set of objects was selected that varied in one, two, or three feature dimensions from trial to trial. The nonvarying features were a spherical body shape, straight arms with blunt ending, gray color, or uniform surface. These feature values were never associated with reward during the experiment and thus represent reward-neutral features. These neutral features defined a neutral object to which we added either one, two, or three nonneutral feature values rendering them 1-D, 2-D, and 3-D, respectively (Figure 1C). For blocks with objects that varied in one feature dimension (1-D attentional load condition), three feature values from that dimension were chosen at the beginning of the block (e.g., body shapes that were oblong, pyramidal, and cubic). One of these features was associated with reward, whereas the two remaining features were not reward associated and thus served as distracting features. Within individual trials, objects never had the same feature values for these dimensions as illustrated for three

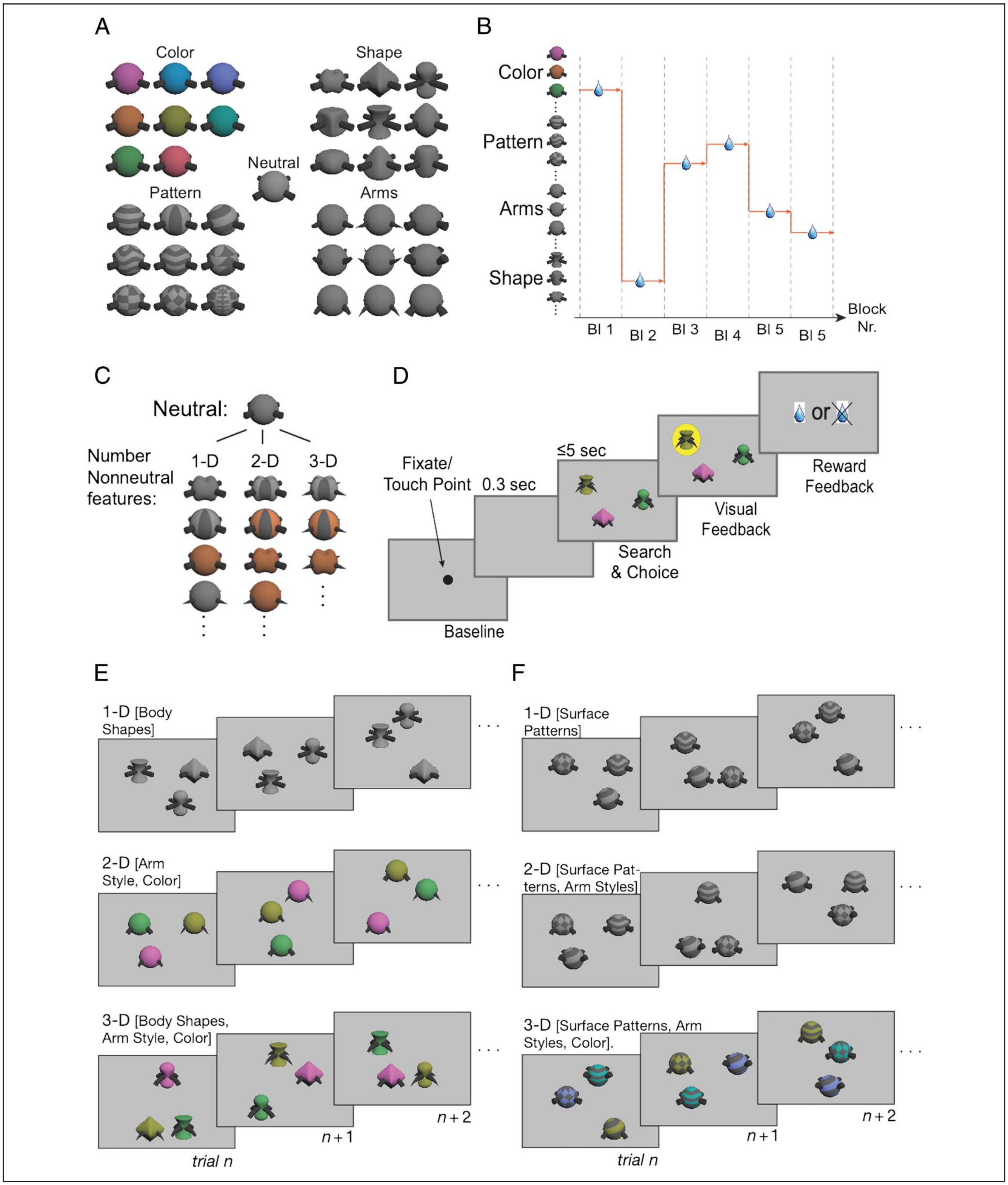


Figure 1. Task paradigm and feature space. (A) The task used 3-D rendered Quaddle objects that varied in color, pattern, shape, and arm style. The features gray color, straight arms, and spherical body shape were never rewarded in any of the experiments and therefore constitute "neutral" features. (B) For successive blocks of 35–60 trials, a single feature was rewarded. (C) The attentional load conditions differed in the number of nonneutral feature dimensions that varied across trials in a block. Blocks with 1-D, 2-D, and 3-D objects contained stimuli varying features in one, two, and three feature dimensions. (D) Trials were initiated by touching or fixating a central stimulus. Three objects were shown at random locations, and subjects had to choose one by either touching (four monkeys) or fixating (two monkeys) an object for ≥0.7 sec. Visual feedback indicated correct (yellow) versus error (gray; not shown) outcomes. Fluid reward followed correct outcomes. (E) Sequences of three example trials for a block with 1-D objects (top row; shape varied), 2-D objects (center; color and arms varied), and 3-D objects (bottom row, body shape, arms, and color). (F) Same as E but for an object set varying surface pattern, arms, and color. Bl = block; Nr. = number.

successive example trials in Figure 1E and F (top row). The feature values of the unused dimensions were the features of the neutral objects in all trials of that block. For blocks with objects varying in two feature dimensions, a set of three feature values per dimension was selected to obtain nine unique objects combining these features. Only one of the features was associated with reward, whereas the other two feature values of that dimension and the feature values of the other dimension were not linked to reward. Figure 1E and F (center row) illustrates three example trials of these blocks of the 2-D attentional load condition. For blocks with objects varying in three feature dimensions (3-D attentional load condition), three feature values per dimensions were selected so that the three presented objects had always different features of that dimension, which led to 27 unique objects combining these features. Again, only one feature was associated with reward in a given block, whereas all other feature values were not linked to reward.

Blocks with objects that varied in one, two, and three feature dimensions constitute 1-D, 2-D, and 3-D attentional load conditions because they vary the number of feature dimensions that define the search space when learning which feature is rewarded. The specific dimension, feature value, and dimensionality of the learning problem varied pseudorandomly from block to block. During individual experimental sessions, monkeys performed up to 30 learning blocks.

Gaze Control

For two animals, gaze was monitored with a Tobii Spectrum with a 600-Hz sampling rate and a binocular infrared eye tracker. For these animals, the experimental session began with a 9-point eye-tracker calibration routine and later reconstruction of object fixations using a robust gaze classification algorithm described in Voloh, Watson, König, and Womelsdorf (2020).

Statistical Analysis

All analyses were performed with custom MATLAB code (Mathworks, Inc.). Significance tests control for the false discovery rate (FDR) with an alpha value of .05 to account for multiple comparisons (Benjamini & Hochberg, 1995).

General Formulation of Rescorla–Wagner RL Models

The value of feature i in trial t, before the outcome was known, is denoted by $V^F_{i,t}$. The superscript F stands for feature, to distinguish it from the value of an object that will be introduced in the next section. The new value $V^F_{i,t+1}$ available for decisions on the next trial depends on which features were at trial t present in the chosen object and whether this choice was rewarded $R_t = 1$, or not $R_t = 0$. The values of features that were present in objects and that were not chosen, as well as those that could appear in the course of the session but were not present on the current trial, decay with a parameter value ω^{RL}_t, where "RL" denotes that the decay component is from the reinforcement component of the model as opposed to the decay of the WM component introduced below. The features that were present in the chosen and rewarded object increase in value, because the reward prediction error (PE), $R_t - V^F_{i,t}$, is positive, whereas when the chosen object was not rewarded, the value decays. We have summarized these update rules in the following equations:

$$V^F_{i,t+1} = V^F_{i,t} + \eta_t\, f^{A,V}_{i,t} \times \left(R_t - V^F_{i,t}\right), \qquad \text{features of chosen objects} \tag{1}$$

$$= \left(1 - \omega^{\mathrm{RL}}_{\mathrm{nc}}\right) V^F_{i,t} \qquad \text{features of nonchosen objects} \tag{2}$$

$$= \left(1 - \omega^{\mathrm{RL}}_{\mathrm{np}}\right) V^F_{i,t} \qquad \text{nonpresented features} \tag{3}$$

The factor $f^{A,V}_{i,t}$ is explained further down. We have indicated a trial dependence in gain η and allow the decay parameter ω to depend on whether the feature was present in the nonchosen object (nc) or whether it was part of the stimulus set of the session but not presented (np) in the current trial. It further carries a superscript RL to indicate it is part of the RL formulation rather than WM (superscript WM). The setting of these parameters depends on the specific model version. In the base RL model, there is no feature-value decay $\omega_{\mathrm{nc},t} = \omega_{\mathrm{np},t} = 0$ and the gain is constant and equal to η. In the next model "RL gain and loss," the gain depends on whether the choice was rewarded (gain) or not rewarded (loss), $\eta_t = \eta_{\mathrm{Gain}}\, R_t + \eta_{\mathrm{Loss}}\, (1 - R_t)$, which introduces two new parameters η_{Gain} and η_{Loss} for rewarded and nonrewarded choices, respectively.

In most models, the decay for nonchosen and not-presented features was equal $\omega_{\mathrm{nc},t} = \omega_{\mathrm{np},t} = \omega^{\mathrm{RL}}$, introducing only a single additional parameter. In the so-called hybrid models, we add a feature-dimension gain factor $f^{A,V}_{i,t}$, which reflects attention to a particular dimension. It is calculated using a Bayesian model (see below) and is indicated by a superscript V because it affects the value update. Hence, it acts as if information about the role of a certain dimension in the acquisition of the reward is not available. The choice probability $p^{\mathrm{RL}}_{i,t}$ for object i at trial t is determined using a softmax function:

$$p^{\mathrm{RL}}_{i,t} = \frac{\exp\left(\beta_t \sum_{j \in O_i} f^{A,CP}_{j,t} V^F_{j,t}\right)}{\sum_k \exp\left(\beta_t \sum_{j \in O_k} f^{A,CP}_{j,t} V^F_{j,t}\right)} \tag{4}$$

The sum in the exponent of the preceding expression is over the features j that are part of object i, which defines the set O_i. The factor β_t in the exponent determines the extent to which the subject exploits, that is, systematically chooses the object with the highest compound value (reflected in a large β), or explores, that is, makes choices independent of the compound value (reflected in small β). In most model versions, the β did not change over trials (parameter β^{RL}), whereas in the meta-learning models with adaptive exploration, its value was adaptive, reflecting the history of reward outcomes, and is thus trial dependent (see the following subsection). The factor $f^{A,CP}_{i,t}$ is a feature-dimension gain factor that acts only in the choice probability; it reflects that some dimensions do not contribute to the choice probability when they are deemed irrelevant (not attended).

Adaptive Exploration

For models with adaptive β_t values, we follow the model of Khamassi, Enel, Dominey, and Procyk (2013), which involves determining an error trace:

$$\beta^*_{t+1} = \beta^*_t + \alpha_+ \max(\delta_t, 0) + \alpha_- \min(\delta_t, 0) \qquad (5)$$

where the min and max functions are used to select the negative and positive parts, respectively, of an estimate of the reward PE,

$$\delta_t = R_t - \frac{1}{\#(j \in O_i)} \sum\nolimits_{j\in O_i} V^F_{j,t} \qquad (6)$$

This is a different form of the PE than above, because here we need to consider all features in a chosen object, rather than each feature separately. The error trace is translated into an actual β value using

$$\beta_t = \frac{\beta_m}{1 + \exp\left(-\omega_1\left(\beta^*_t - \omega_2\right)\right)} \qquad (7)$$

This adaptive component replaces one parameter by five new parameters: α_+, α_-, β_m, ω_1, and ω_2, of which we fixed four in most models to the following values that came out of pilot parameter explorations, namely, $\alpha_+ = -0.6$, $\alpha_- = -0.4$, $\omega_1 = -6$, and $\omega_2 = 0.5$, and varied β_m and sometimes varied α_- as well.

Attentional Dimension Weight

The attentional gain factor ($f^{A,V}_{i,t}$ and $f^{A,CP}_{i,t}$) uses a Bayesian estimate of what the target feature f is, hence what the relevant feature dimension is, and weighs the contribution of each feature value according to whether it is part of the target dimension (Oemisch et al., 2019; Hassani et al., 2017; Niv et al., 2015). From the target feature probability $p(f|\mathcal{D}_{1:t})$ (see Equations 5–7 in Hassani et al. [2017] for the derivation of the equation we use to update this probability from trial to trial), we can obtain a target dimension probability by summing over all the feature values $f(d)$ that belong to a particular dimension d,

$$p^D_{d,t} = p(d|\mathcal{D}_{1:t}) = \sum\nolimits_{f\in f(d)} p(f|\mathcal{D}_{1:t}) \qquad (8)$$

this is turned into a feature gain

$$\phi^A_{d,t} = \frac{\left(p^D_{d,t}\right)^{\alpha}}{\sum_e \left(p^D_{e,t}\right)^{\alpha}} \qquad (9)$$

which weighs feature values in each object according to their dimension $d(f)$; for an object i, this becomes $V_{i,t} = \sum_{j\in O_i} \phi^A_{d(j),t} V^F_{j,t}$, and which we incorporate as a feature-dependent factor $f^A_{i,t} = \phi^A_{d(i),t}$ in the relevant expressions (Equations 1 and 4, for the value and choice probability, indicated with additional superscripts V and CP, respectively).

Stickiness in Choice Probability

Stickiness of choosing objects refers to choosing the object whose feature values overlap with the previously chosen one and represents perseveration (Balcarras, Ardid, Kaping, Everling, & Womelsdorf, 2016). It is implemented by making the choice probability dependent on whether a feature on the previous trial is present in an object.

$$p^S_i = \frac{\exp\left(\beta_t \sum_{j\in O_i} f^{A,CP}_{j,t} V^F_{j,t}\right) + \Delta_{t-1,i}}{\sum_k \exp\left(\beta_t \sum_{j\in O_k} f^{A,CP}_{j,t} V^F_{j,t}\right) + \Delta_{t-1,i}} \qquad (10)$$

Here, $\Delta_{t-1,i}$ is equal to $e^{\gamma} - 1$ when object i presented on trial t contains at least one feature that was also present in the chosen object on the previous trial ($t - 1$). By subtracting 1, we ensure that when $\gamma = 0$, there is no stickiness contribution to the choice. In our setup, it is possible that more than one of the current objects contain features that were present in the previously chosen object.

Combined WM/RL Models

WM models are formulated in terms of the value $V^{\mathrm{WM}}_{i,t}$ of an object i irrespective of what features are present in it (Collins & Frank, 2012). These values are initialized to a noninformative value of $\frac{1}{n_o}$, where n_o is the number of objects. When each of the objects has this value, there is no preference in choosing one above the other. When an object is chosen on trial t, the value is set to $V^{\mathrm{WM}}_{i,t+1} = 1$ when rewarded, whereas it is reset to the original value $V^{\mathrm{WM}}_{i,t+1} = \frac{1}{n_o}$ when the choice was not rewarded. All other values decay toward the original value with a decay parameter ω^{WM}:

$$V^{\mathrm{WM}}_{i,t+1} = V^{\mathrm{WM}}_{i,t} - \omega^{\mathrm{WM}}\left(V^{\mathrm{WM}}_{i,t} - \frac{1}{n_o}\right) \qquad (11)$$

The values are then directly used in the choice probabilities (also denoted p_{Choice}):

$$p^{\mathrm{WM}}_{i,t} = \frac{\exp\left(\beta^{\mathrm{WM}} V^{\mathrm{WM}}_{i,t}\right)}{\sum_j \exp\left(\beta^{\mathrm{WM}} V^{\mathrm{WM}}_{j,t}\right)} \qquad (12)$$

This component mechanism thus introduces two new parameters, a decay parameter ω^{WM} and the softmax parameter β^{WM}, which are separately varied in the fitting procedure.

Integrating Choice Probabilities

In the most comprehensive models, choices are determined by a weighted combination of the choice probabilities derived from the RL and WM components, referred to as $p_{i,t}^{T}$ (T stands for total),

$$p_{i,t}^{T} = w_t p_{i,t}^{\mathrm{WM}} + (1 - w_t) p_{i,t}^{\mathrm{RL}} \tag{13}$$

A larger w_t means more weight for the WM predictions in the total choice probability. The update of w_t reflects the value of the choice probability for the choice made and the capacity limitations of the WM:

$$w_{t+1} = \frac{p_t^{\mathrm{WMC}} w_t}{w_t p_t^{\mathrm{WMC}} + (1 - w_t) p_t^{\mathrm{RLC}}} \tag{14}$$

where

$$p_t^{\mathrm{RLC}} = p_{a(t),t}^{\mathrm{RL}}\ r_t + \left(1 - p_{a(t),t}^{\mathrm{RL}}\right)(1 - r_t).$$

This expression selects from among two possible values for p_t^{RLC} depending on whether $r_t = 1$ or 0. Here, $a(t)$ is the index of the object chosen on trial t. In addition,

$$p_t^{\mathrm{WMC}} = \alpha\left(p_{a(t),t}^{\mathrm{WM}}\ r_t + \left(1 - p_{a(t),t}^{\mathrm{WM}}\right)(1 - r_t)\right) + (1 - \alpha)\left(\frac{1}{n_{\mathrm{O}}}\right) \tag{15}$$

where $\alpha = \min(1, \frac{C_{\mathrm{WM}}}{n_{\mathrm{S}}})$ and C_{WM} is the WM capacity, essentially the number of objects about which information can be accessed, and n_{S} is the number of objects that can be presented during the task. It is determined as the number of objects whose value $V_{i,t}^{\mathrm{WM}}$ exceeds $\frac{1}{n_{\mathrm{o}}}$ by a margin of 0.01. When n_{S} is much larger than C_{WM}, the information in p^{WM}, which is unlimited in capacity but decays with time, cannot be read out, instead $p_t^{\mathrm{WMC}} = 1/n_{\mathrm{O}}$. Hence, when $p_{a(t),t}^{\mathrm{RL}}$ exceeds $\frac{1}{n_{\mathrm{o}}}$, it will win the competition for influence and reduce w_{t} toward zero and, with that, the influence of WM via $p_{i,t}^{\mathrm{WM}}$.

Posterior Predictive Checks and Model Identifiability

Our overall focus was to evaluate how fit parameters vary with task condition and across subjects. For this, it is necessary that the parameters have a clear meaning, that they are reproducible (model identifiability), and that the objective function value correlates with the degree to which model choices match the subject's choices (a kind of posterior predictive test). To assess this, we performed a number of validation analyses on the model that we had found to best fit the subject's choices (the top-ranked model in Figure 3A, Model 1; Table 1), which was characterized by eight fitting parameters, which could present a challenge to fitting procedures.

Our objective function was the negative log likelihood (NLL), which was minimized through a call to the MATLAB function *fminsearch* followed by a call to *fmincon*. We did this for multiple different initial conditions and found a few distinct solutions, each converged to from multiple different initial values and each corresponding to slightly different values of the objective function. These differences were typically smaller than the differences between different models. This shows that the algorithm can get stuck in local minima. Note that this did not occur for models with fewer parameters.

We evaluated the behavioral performance of each of the corresponding parameter values by generating choice sequences based on sampling randomly according to the model-generated choice probabilities, repeating the sampling multiple times for each unique parameter set derived from an initial condition (see Appendix; Appendix Figure A1). The posterior predictive performance was characterized by determining the mean reward and the overlap between the model's and the subject's choices—quantified as the fraction of common choices. These performance measures are different from the NLL value that we minimized, because there we took the subject's choices and the received rewards to update the (feature) values in the model and determined the choice probability for each trial according to the model; the negative log of these adds to the objective function. In contrast, here we generate choices, often different from those of the subject because we use the choice probabilities from the model for the fitted parameter values but updated trial-to-trial with model choices and the corresponding rewards according to the task rule. The resulting choice–reward sequence generated is also different from the measured one, but it is the one used to update the model's feature values across trials. It is therefore likely to generate different choices across blocks than the subject even when using the same stimulus sequence. The model can be considered good when the experimentally generated sequence cannot be distinguished from the distribution of the model-generated choices; this can occur for rather low values of overlap, when the choice probabilities are much smaller than 1.

We find that the lower the NLL, the higher the overlap in choices is (Appendix Figure A1A). This means that NLL is a useful indicator of the quality of the fit. For the exploratory model runs, the overlap in choices is low but does exceed 0.5. When considering the same distribution of choices across options, but made randomly, which we accomplished by randomly permuting the model's choices, the overlap is reduced by about 0.04 (8%; Appendix Figure A1B), ending up below 0.5. Note that this is higher than expected based on purely random

Table 1. Overview of the Parameter Used in Models Evaluated and Ranked According to BIC Higher than the Base RL Model (Which is the Model Ranked 26th)

Model Rank	*RL Gain* (η)	*RL Decay* (ω)	*RL Softmax* (β)	*RL Atn* (α)	*WM Decay* (ω)	*WM Softmax* (β)	*WM Capacity* (C_{WM})	*Stickiness* (γ)	*#para*
1	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	ω^{RL}	β_m, α_- (α_+, ω_1, ω_2 fix)	–	ω^{WM}	β^{WM}	C_{WM}	–	8
2	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	ω^{RL}	β_m (α_+, α_-, ω_1, ω_2 fix)	–	ω^{WM}	β^{WM}	C_{WM}	γ	8
3	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	ω^{RL}	β^{RL}	–	ω^{WM}	β^{WM}	C_{WM}	γ	8
4	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	ω^{RL}	β_m (α_+, α_-, ω_1, ω_2 fix)	–	ω^{WM}	β^{WM}	C_{WM}	–	7
5	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	ω^{RL}	β^{RL}	–	ω^{WM}	β^{WM}	C_{WM}	–	7
6	η	ω^{RL}	β_m (α_+, α_-, ω_1, ω_2 fix)	–	ω^{WM}	β^{WM}	C_{WM}	–	7
7	η	ω^{RL}	β^{RL}	–	ω^{WM}	β^{WM}	C_{WM}	–	6
8	η	ω^{RL}	β_m (α_+, α_-, ω_1, ω_2 fix)	–	ω^{WM}	β^{WM}	C_{WM}	–	6
9	η	ω^{RL}	β_m, α_- (α_+, ω_1, ω_2 fix)	–	ω^{WM}	β^{WM}	C_{WM}	γ	7
10	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	ω^{RL}	β_m, α_- (α_+, ω_1, ω_2 fix)	–	–	–	–	–	5
11	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	ω^{RL}	β_m (α_+, α_-, ω_1, ω_2 fix)	–	–	–	–	γ	5
12	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	ω^{RL}	β^{RL}	–	–	–	–	γ	5
13	η	ω^{RL}	β_m (α_+, α_-, ω_1, ω_2 fix)	–	–	–	–	–	3
14	η	ω^{RL}	β^{RL}	–	–	–	–	γ	4
15	η	ω^{RL}	β^{RL}	–	–	–	–	–	3
16	η	ω^{RL}	β^{RL}	α	ω^{WM}	β^{WM}	C_{WM}	–	7
17	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	ω^{RL}	β^{RL}	–				–	4
18	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	ω^{RL}	β^{RL}	α	ω^{WM}	β^{WM}	C_{WM}	–	8
19	η	–	β_m (α_+, α_-, ω_1, ω_2 fix)	–	ω^{WM}	β^{WM}	C_{WM}	–	5
20	η	ω^{RL}	β^{RL}	α	ω^{WM}	β^{WM}	C_{WM}	γ	8
21	η	–	β^{RL}	–	ω^{WM}	β^{WM}	C_{WM}	–	5
22	η	–	β^{RL}	–	ω^{WM}	β^{WM}	C_{WM}	γ	6
23	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	–	β^{RL}	–	ω^{WM}	β^{WM}	C_{WM}	–	6
24	η	–	β^{RL}	–	–	–	–	γ	3
25	$\eta_{\text{Gain}}, \eta_{\text{Loss}}$	–	β^{RL}	–	–	–	–	–	3
26	η	–	β^{RL}	–	–	–	–	–	2

See Figure 3 text for a graphical illustration of the model rank ordering. Atn = Attention.

choices with equal probability for each option, because both the subject and the model make the correct choice more than chance and we labeled the correct choice as Option 1. The upshot is that lower NLL leads to more overlap in choices and therefore is a good basis to compare models with.

A typical issue for fitting functions with many parameters is shallow optima in which similar values for the objective function are found when covarying two or more parameters. This could affect the identifiability of the model, because a given parameter setting would generate model choices that would be most optimally fit by quite different parameter values. As mentioned in the preceding text, we ran the optimization procedure for multiple initial conditions and then generated multiple-choice sequences for each of them. We fitted each of these sequences again. This gave us a multidimensional distribution of parameters' values (Appendix Figure A1C). We found that three parameters displayed a larger range of values than the rest of them. Specifically, for some initial conditions, they created outlier values that were close to the upper-bound constraints we imposed on the optimization with *fmincon*; these outlier values were also present in the fits of the model-generated sequences. The most affected was the beta parameter for WM (β^{WM}), as well the WM capacity (C_{WM}), and to a lesser extent, also the (adaptive) beta parameter for RL model (β_{m}). The underlying cause was the covariation between C_{WM} and β^{WM}, which we observed as a clear correlation between two fitting parameters across the refitted parameter values (Appendix Figure A1D). We address this effect by using a cross-validation procedure (see below) to deal with the outlier parameter values.

Model Comparison and Cross-validation

To compare models, we calculated the log likelihood (normalized by the number of choices) of each model fit to the choices of the monkeys and computed the Bayesian information criterion (BIC) for each model that penalizes models according to the number of free parameters. We rank ordered the BIC to identify the model most predictive of the monkey's choices. We used a cross-validation procedure for validating that the best-fit models do not overfit the data. For the cross-validation, we evaluated how well a model predicted in terms of the NLL the subject's choices of (test) learning blocks that were withheld when fitting the model parameters on the remaining (training) data sets. We repeated the cross-validation 50 times and used the average parameter values across these 50 cross-validation runs to simulate the choices of the monkey. For each cross-validation, we cut the entire data set at two randomly chosen blocks, yielding three parts. The two largest parts were assigned as training and test sets. We did this to keep the trials in the same order as the monkey performed them, as the memory-dependent effects in the model (and presumably the monkey) extend beyond the block boundaries. This is different from the standard procedure, where blocks were randomly assigned to test and training sets, hence breaking the block ordering that is important for the model. In general, the cross-validation results were qualitatively similar to the results optimizing the entire data set and gave near-identical rank ordering of the models with identical top-ranked models. This finding rules out overfitting.

Relation of Model Parameter Values and Behavior

To test how each of the model parameters of the best-fitting model related to the learning and performance levels across different attentional load conditions, we constructed linear mixed effects (LME) models (Pinherio & Bates, 1996). The models predict the learning speed *LS* (corresponding to the number of trials to criterion performance) or the plateau accuracy *AC* (percent correct over trials after criterion was reached) based on the individual model parameter values (Par1, Par2, … Par$_n$) and the attentional load condition (with three levels for the 1-D, 2-D, and 3-D load conditions). All models used as random effects the factor Monkeys (each of six animals) to control for individual variations. For the best-fitting model, this LME had the form:

$$\begin{aligned} \mathit{LS\ or\ Accuracy} = {} & Par1_{\beta^{\mathrm{WM}}} + Par2_{C_{\mathrm{WM}}} \\ & + Par3_{\omega^{\mathrm{WM}}\mathrm{WM\ decay}} \\ & + Par4_{\beta_m\,*\,\mathrm{adaptive\ exploration}} \\ & + Par5_{\alpha+\mathrm{amplitude\ for\ adaptive\ exploration}} \\ & + Par6_{\omega^{\mathrm{RL}}} + Par7_{\eta_{\mathrm{Gain}}} + Par8_{\eta_{\mathrm{loss}}} \\ & + Att_{\mathrm{Load}} + (1|Monkey) + b + \varepsilon \end{aligned} \tag{16}$$

In a second LME analysis, we tested which learning parameter values of the best-fit model are able to predict fast versus slow learners. To test this, we ranked subjects by their learning speed (their average trials-to-criterion) for each attentional load condition. We tested whether the rank ordering (learner rank) was accounted for by individual model parameter values. Using the parameter values of the best-fitting model, we tested the LME of the form:

$$\begin{aligned} \mathit{Learner}\ \mathrm{rank} = {} & Par1_{\beta^{\mathrm{WM}}} + Par2_{C_{\mathrm{WM}}} \\ & + Par3_{\omega^{\mathrm{WM}}\mathrm{WM\ decay}} + Par4_{\beta_m\,*\,\mathrm{adapt.\ explor.}} \\ & + Par5_{\alpha+\mathrm{ampl.\ for\ adapt.\ explor.}} + Par6_{\omega^{\mathrm{RL}}} \\ & + Par7_{\eta_{\mathrm{Gain}}} + Par8_{\eta_{\mathrm{loss}}} + (1|Att_{\mathrm{Load}}) \\ & + b + \varepsilon \end{aligned} \tag{17}$$

All inference statistics of the linear effects models account for multiple comparisons by adjusting the significance level according to an FDR of $p = .05$.

RESULTS

Behavioral Performance

We measured how six monkeys learned the relevance of object features in a learning task while varying the number of reward-irrelevant, distracting feature dimensions of these objects from one to three. On each trial, subjects chose one of three objects and either did or did not receive reward, to learn by trial-and-error which object feature predicted reward. The rewarded feature could be any one of 37 possible features values from four different feature dimensions (color, shape, pattern, and arms) of multidimensional Quaddle objects (Figure 1A; Watson, Voloh, Naghizadeh, et al., 2019). The rewarded feature, that is, the reward rule, stayed constant within blocks of 35–60 trials (Figure 1B). Learning blocks varied in the number of nonrewarded, distracting features (Figure 1C). Subjects had 5 sec to choose an object that triggered correct or error feedback in the form of a yellow or gray halo around the chosen object, respectively (Figure 1D). The first of three experimental conditions was labeled 1-D attentional load because all the distractor features were from the same dimension as the target feature (e.g., different body shapes; see examples in the top row of Figure 1E and F). At 2-D attentional load, features of a second dimension varied in addition to features from the target feature dimension (e.g., objects varied in body shapes and surface patterning). At 3-D attentional load, object features varied along three dimensions (e.g., varying in body shapes, surface patterns, and arm styles; bottom row in Figure 1E and F).

Six monkeys performed a total number of 989 learning blocks, completing on average 55/56/54 (*SE* = 4.4/4.3/4.2, range = 41–72) learning blocks for the 1-D, 2-D, and 3-D attentional load conditions, respectively. The number of trials in a block needed to learn the relevant feature, that is, to reach 75% criterion performance increased for the 1-D, 2-D, and 3-D attentional load condition from, on average, 6.5, 13.5, and 20.8 trials (*SE*s = 4.2/8.3/6.9; Kruskal–Wallis test, $p = .0152$; ranks: 4.8, 10.2, and 13.6; Figure 2A and B). Learning speed did not differ when the rewarded feature in a block was of the same or of a different dimension as the rewarded feature in the immediately preceding block (intradimensional vs. extradimensional block transitions; Wilcoxon rank sum test, $p = .699$, rank sum = 36; Figure 2C).

Flexible learning can be influenced by target- and distractor-history effects (Banaie Boroujeni, Watson, & Womelsdorf, 2020; Rusz, Le Pelley, Kompier, Mait, & Bijleveld, 2020; Chelazzi, Marini, Pascucci, & Turatto, 2019; Failing & Theeuwes, 2018; Le Pelley, Pearson, Griffiths, & Beesley, 2015), which may vary with attentional load. We tested this by first evaluating the presence of latent inhibition, which refers to slower learning of a newly rewarded target feature when that feature was a (learned) distractor in the preceding block compared to when the target feature was not shown in the previous block. We did, however, not find a latent inhibition effect (paired signed rank test: $p = .156$, signed rank = 3; Figure 2D, left). A second history effect is persevering choosing the feature that was a target in the previous block. We quantified this target perseveration by comparing learning in blocks in which a previous (learned) target feature became a distractor, to learning blocks in which distractor features were new. We found that target perseveration significantly slowed down learning (paired signed rank test: $p = .0312$, signed rank = 0; Figure 2D, right), which was significantly more pronounced in the high (3-D) than in the low (1-D) attentional load condition (paired signed rank test, again: $p = .0312$, signed rank = 0; Figure 2E). These learning history effects suggest that learned target features had a significant influence on future learning in our task, particularly at high attentional loads, whereas learned distractors had only marginal or no effects on subsequent learning.

Multicomponent Modeling of Flexible Learning of Feature Values

To discern specific mechanisms underlying flexible feature-value learning in individual monkeys, we fit a series of RL models to their behavioral choices (see Methods). These models formalize individual cognitive learning mechanisms and allowed characterizing their role in accounting for behavioral learning at varying attentional loads. We started with the classical Rescorla–Wagner reinforcement learner that uses two key mechanisms: (i) the updating of value expectations of features V^F every trial t by weighting reward PEs with a learning gain η: $V^F_{i,t+1} = V^F_{i,t} + \eta(R_t - V^F_{i,t})$ (with reward $R_t = 1$ for a rewarded choice and zero otherwise) and (ii) the probabilistic ("softmax") choice of an object O given the sum of the expected values of its constituent features V_i, $p\text{Choice}^{\text{RL}}_i = \frac{\exp\left(\beta^{\text{RL}}\sum_{j\in O_i} V^F_j\right)}{\sum_j \exp\left(\beta^{\text{RL}}\sum_{k\in O_j} V^F_k\right)}$ (Sutton & Barto, 2018). These two mechanisms incorporate two learning parameters: the weighting of PE information by η (often called the learning rate), $\eta(PE)$, and the degree to which subjects explore or exploit learned values represented by β^{RL}, which is small or close to zero when exploring values and larger when exploiting values.

We augmented the Rescorla–Wagner learning model with up to seven additional mechanisms to predict the monkey choices (Table 1; see Discussion and Appendix for the results with non-Rescorla–Wagner models such as attentional switching and hypothesis-testing models). The first of these mechanisms enhanced the expected values of all object features that were chosen by decaying feature values of nonchosen objects. This selective decay improved the prediction of choices in reversal learning and probabilistic multidimensional feature learning tasks (Oemisch et al., 2019; Hassani et al., 2017; Radulescu, Daniel, & Niv, 2016; Niv et al., 2015; Wilson & Niv,

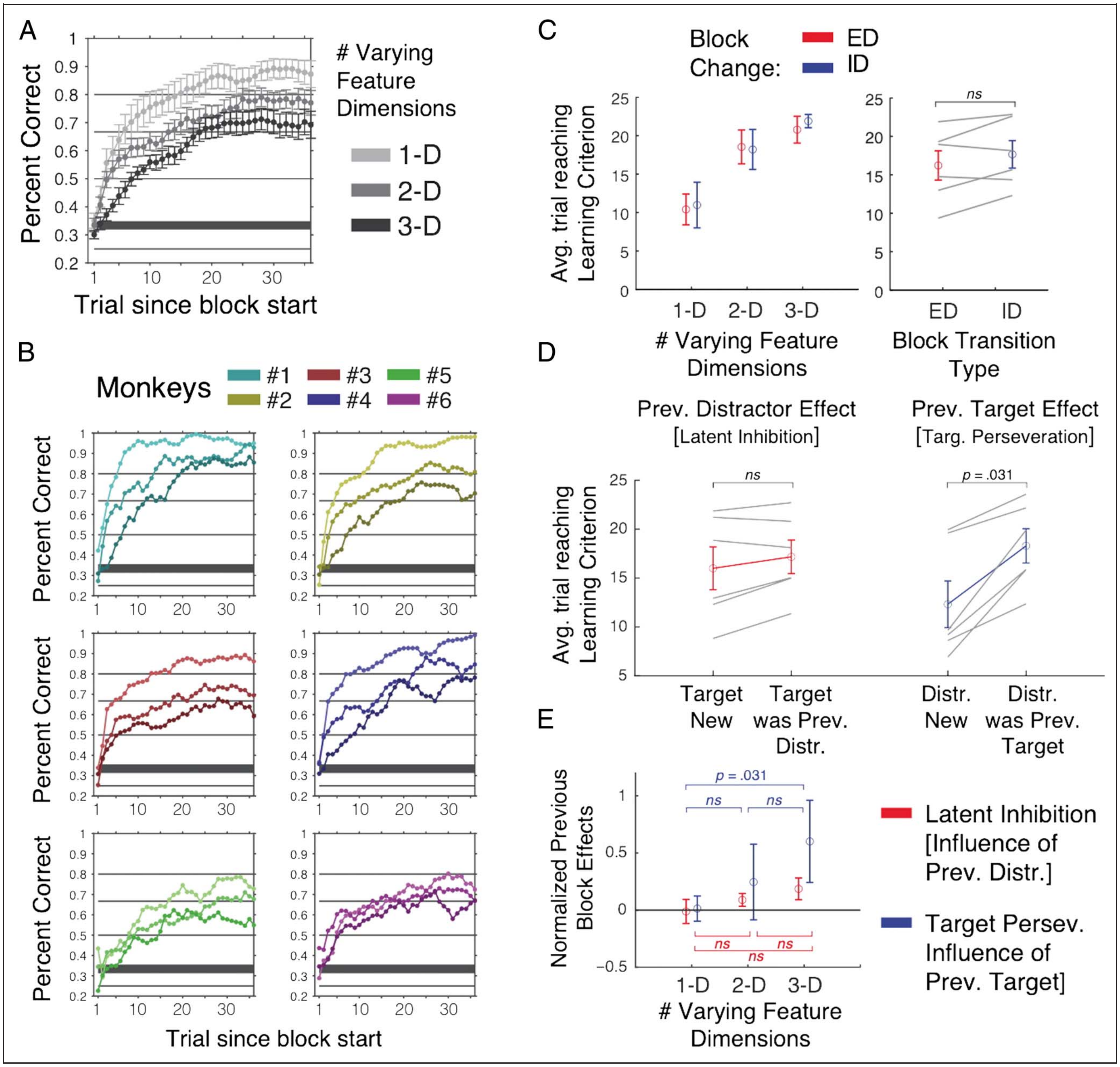


Figure 2. Learning performance. (A) Average learning curves across six monkeys for the 1-D, 2-D, and 3-D load conditions. (B) Learning curves for each monkey (colors) for 1-D, 2-D, and 3-D (low-to-high color saturation levels). All monkeys showed fastest learning for the low-load condition and slowest learning for the high-load condition. Curves are smoothed with a five-trial forward-looking window. (C, left) The average trials-to-criterion (75% accuracy over 10 consecutive trials) for low to high attentional loads (*x* axis) for blocks in which the target feature was either of the same (intradimensional [ID]) or different (extradimensional [ED]) dimension—as in the preceding trial. (C, right) Average number of trials-to-criterion across load conditions. Gray lines denote individual monkeys. Errors are *SE*. (D) The red color denotes average trials-to-criterion for blocks in which the target feature was novel (not shown in the previous block) or when it was previously a learned distractor. The blue color denotes the condition in which a distractor feature was either novel (not shown in the previous block) or part of the target in the previous block. When distractors were previously targets, learning was slower. (E) Latent inhibition of distractors (red) and target perseveration (blue) at low, medium, and high loads. Errors indicate *SE*. Avg. = average; Dist. = distractor; Persev. = perseveration; Prev. = previous.

2011). It is implemented as decay ω^{RL} of feature values V_i from nonchosen features and thereby enhanced the value estimate for chosen (and hence attended) features for the next trial t:

$$V_{i,t+1}^{F} = \left(1-\omega^{\mathrm{RL}}\right)V_{i,t}^{F} \quad \text{(Decay of nonchosen feature values)} \tag{18}$$

As a second mechanism, we considered a WM process that uploads the identity of rewarded objects in an STM. Such a WM can improve learning of multiple stimulus–response mappings (Collins, Ciullo, Frank, & Badre, 2017; Collins & Frank, 2012) and multiple reward locations (Viejo et al., 2018; Viejo, Khamassi, Brovelli, & Girard, 2015). Similar to Collins and Frank (2012), we uploaded the value of an object in WM (V_i^{WM}) when it was chosen and rewarded and decayed its value with a time constant $\frac{1}{\omega^{WM}}$. WM proposes a choice using its own choice probability $p\text{Choice}^{WM}$, which competes with the $p\text{Choice}^{RL}$ from the RL component of the model. The actual behavioral choice is the weighted sum of the choice probabilities of the WM and RL components $w(p\text{Choice}^{WM}) + (1 - w)p\text{Choice}^{RL}$. A weight of $w > 0.5$ would reflect that the WM content dominates the choice, which would be the case when the WM capacity can maintain object values for sufficiently many objects before they fade away (see Methods). This WM module reflects a fast "one-shot" learning mechanism for choosing the recently rewarded object.

As a third mechanism, we implemented a meta-learning process that adaptively increases the rate of exploration (the β parameter of the standard RL formulation) when errors accumulate. Similar to Khamassi et al. (2013), the mechanism uses an error trace β_t^*, which increases when a choice was not rewarded, by an amount proportional to the negative PE for that choice with a negative gain parameter α_-, and decreases after correct trials proportional to the positive PE weighted by a positive gain parameter α_+ (Khamassi et al., 2013):

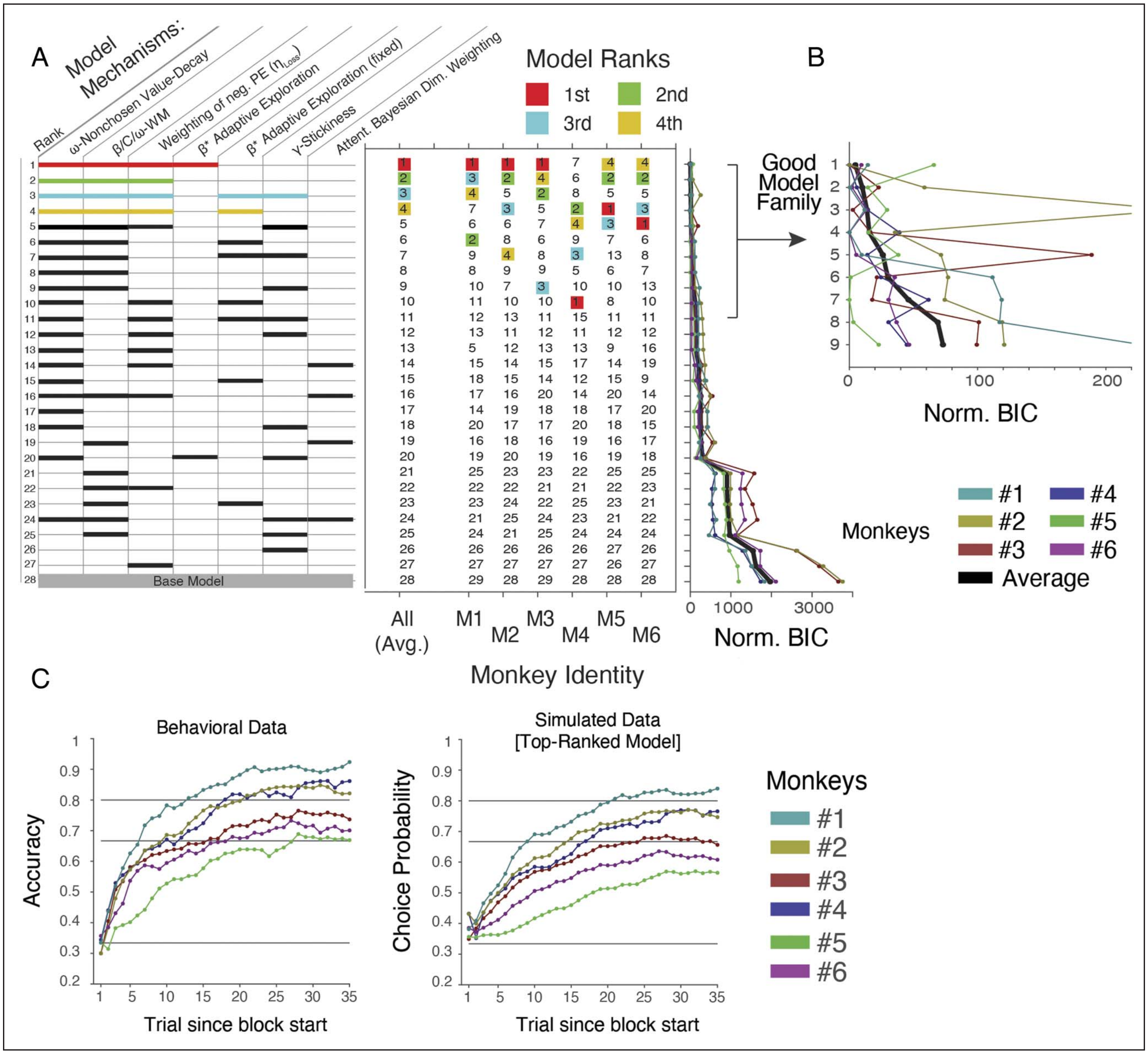
$$\beta_{t+1}^* = \beta_t^* + \alpha_+[\delta_t]_+ - \alpha_-[-\delta_t] \qquad \text{(Adjustment of exploration rate)} \tag{19}$$

where the PE is given by $\delta_t = R_t - V_t$, with V reflecting the mean of all the feature values of the chosen object. The error trace contains a record of the recent reward performance and was transformed into a beta parameter for the softmax choice according to $\beta_t^{RL} = \frac{\beta_{\max}}{1+\exp\left(-\omega_1\left(\beta_t^* - \omega_2\right)\right)}$ (Khamassi et al., 2013). Transiently increasing the exploration rate increases the chances to find relevant object features when there are no reliable, learned values to guide the choice and there are multiple possible feature dimensions that could be valuable. We kept $\alpha_+ = -0.6$, $\alpha_- = -0.4$, $\omega_1 = -6$, and $\omega_2 = 0.5$ fixed and varied $\beta_{\max}$ and, in some cases, α_- as well, resulting in a fourth model mechanism that could underlie flexible feature learning under increasing attentional load.

We tested three other neurobiologically plausible candidate mechanisms that played important roles in prior learning studies. A fifth mechanism implemented choice stickiness to account for perseverative (repetitive) choices independent of value estimates (Balcarras et al., 2016; Badre, Doll, Long, & Frank, 2012). A sixth parameter realized an "attentional" dimension weight during value updates, which is realized by multiplying feature values given the reward likelihood for the feature dimension they belong to (Oemisch et al., 2019; Leong et al., 2017). Finally, as a seventh parameter, we separately modeled the weighting of negative PEs after error outcomes, η_{Loss}, and the weighting of positive PEs for correct outcomes, η_{Gain}, to allow separate learning speeds for avoiding objects that did not lead to reward (after negative feedback) and for facilitating choices to objects that led to rewards (after positive feedback; Taswell, Costa, Murray, & Averbeck, 2018; Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017; Cazé & van der Meer, 2013; van den Bos, Cohen, Kahnt, & Crone, 2012; Kahnt et al., 2009; Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Frank, Seeberger, & O'Reilly, 2004). We constructed models that combined two, three, or four of these mechanisms. This led to models with two to eight free parameters (see Methods). Each model was fitted to the behavioral data separately for each attentional load condition and for each individual monkey. We calculated the BIC to rank order the models according to how well they predicted actual learning behavior given the number of free parameters.

WM, Adaptive Exploration, and Decaying Distractor Values Supplement RL

We found that, across monkeys and attentional load conditions, the RL model that best predicted monkeys' choices during learning had four nonstandard components: (i) WM, (ii) nonchosen value decay, (iii) adaptive exploration rate, and (iv) a separate gain for negative PEs (η_{Loss}; Figure 3). This model had the lowest BIC on average across all six monkeys and was ranked 1st for three individual monkeys (Monkeys 1, 2, and 3; Figure 3A and B; Table 1 shows the complete list of free model parameters for the rank-ordered models). The overall best-ranked model ranked 4th, 5th, and 10th for the other three monkeys (Figure 3A). The top-ranked model for these three monkeys had three of the four mechanisms in common with the overall best-fit model and differed only in one parameter to the overall best-fit model. In other words, the learning performance of monkeys was best fit with a model that incorporated the selective forgetting (feature value decay; for all six monkeys), a separate WM component (with three free parameters; for all six monkeys), an adaptive exploration rate with or without a free parameter (for all six monkeys), and a separate learning rate for error outcomes (for five of six monkeys; Figure 3A). The one monkey (Monkey 4) whose top-ranked model did not use a learning rate for error outcomes had instead the choice stickiness mechanism as part of his best-fit model (the overall seventh model in Figure 3A) and included the learning rate for errors in his fourth best-ranked model (Figure 3A). Together, these findings identify a "family" of good learning mechanisms across monkeys. Within this family, four cognitive learning mechanisms most consistently contributed to predict learning performance. One additional mechanism (choice stickiness) played an important role in one of the six animals but was not needed to account for the behavior of the other five animals. Simulating the monkeys' choices with the overall best-fitting model showed



All (Avg.)	M1	M2	M3	M4	M5	M6
1	1	1	1	7	4	4
2	3	2	4	6	2	2
3	4	5	2	8	5	5
4	7	3	5	2	1	3
5	6	6	7	4	3	1
6	2	8	6	9	7	6
7	9	4	8	3	13	8
8	8	9	9	5	6	7
9	10	7	3	10	10	13
10	11	10	10	1	8	10
11	12	13	11	15	11	11
12	13	11	12	11	12	12
13	5	12	13	13	9	16
14	15	14	15	17	14	19
15	18	15	14	12	15	9
16	17	16	20	14	20	14
17	14	19	18	18	17	20
18	20	17	17	20	18	15
19	16	18	16	19	16	17
20	19	20	19	16	19	18
21	25	23	23	22	25	25
22	22	22	21	21	22	23
23	23	24	22	25	23	21
24	21	25	24	23	21	22
25	24	21	25	24	24	24
26	26	26	26	26	27	26
27	27	27	27	27	26	27
28	29	28	29	28	28	28

Monkey Identity

Figure 3. Rank ordering of models with different combinations of mechanisms. (A) Models (rows) using different combinations of model mechanisms (columns) are rank ordered according to their BIC. The top-ranked model combined four mechanisms that are highlight in red: decay of nonchosen features, WM, adaptive exploration rate, and a separate learning gain for errors (losses). The 2nd, 3rd, and 4th ranked models are denoted with cyan, green, and yellow bars, respectively. Thick horizontal bar indicates that the model mechanism was used in that model. The 26th ranked model was the base RL model that used only a beta softmax parameter and a learning rate. (A, right) Model rank average (first column) and for each individual monkey (Columns 2–7). See Table 1 for the same table in numerical format with additional information about the number of free parameters for each model. (B) After subtracting the BIC of the 1st ranked model, the normalized BICs for each monkey confirm that the top-ranked model has low BIC values for each monkey. (C) Average behavioral learning curves for the individual monkeys (left) and the simulated choice probabilities of the top-ranked model for each monkey. The simulated learning curves are similar to the monkey learning curves providing face validity for the model. Attent. = attentional; Dim. = dimension; neg. = negative; Norm. = normalized.

that it reproduced well the variable learning curves obtained from the monkeys (Figure 3C).

To discern how the individual model mechanisms of the most predictive model contributed to the learning at low, medium, and high attentional loads, we simulated the choice probabilities for this full model as well as for partial models that implemented only individual mechanisms of that full model separately for each load condition (Figure 4A and B). The simulations used the parameter values of the overall best-fit model. This analysis confirmed that the best-fit full model was most closely predicting choices of the animals in all load conditions, showing a difference between the model choice probabilities and the monkeys' choice accuracy of only ~7%

across all three attentional load conditions (Figure 4C). The reduced (partial) model that performed similarly well across all attentional loads used the decay of nonchosen features (ω^{RL}) (ranked 17th among all models; Figures 3A and 4C). All other partial models were performing differently at low and high attentional loads. The partial model having only the WM component (with ω^{WM}) predicted choices well for the 1-D and 2-D load conditions but showed a sharply reduced predictability for choices in the 3-D load condition (Figure 4C). The partial model with the adaptive exploration rate (β^*) worsened choice probability for the low-load condition relative to the standard RL but improved predictions for the 2-D load condition (Figure 4C). Similarly, the partial model with the separate weighting of negative PEs (η_{Loss}, ranked 27th; see Figure 3A) showed overall better choice probabilities than the standard RL model (ranked 28th) but still failed predicting 12%–18% of the monkeys' choices when used as the only nonstandard RL mechanisms (Figure 4C). These results highlight that the selective forgetting of nonchosen values, which is formalized as the decay of nonchosen features (ω^{RL}) was the only parameter that was similarly important across all attentional load conditions. All other cognitive learning mechanisms had functional roles that varied with attentional load.

To understand why WM was only beneficial at low and medium attentional loads but detrimental at high attentional loads, we visualized the choice probabilities that the WM module of the full model generated for different objects. We contrasted these WM choice probabilities with the choice probabilities for different stimuli of the RL module and of the combined WM + RL model (Figure 5A). After a block switch, the WM module uploaded an object as soon as it was rewarded and maintained that rewarded object in memory over only a few trials. When the rewarded object was encountered again before decaying to zero, it guided the choice of that object beyond what the RL module would have suggested (evident in Trial 6 in Figure 5A–C). This WM contribution is beneficial when the same object instance reoccurred within few trials, which happened more frequently with low and medium attentional loads, but only rarely during high loads. At this high-load condition, it was the RL model component that is faithfully tracking the choice probability of the monkey, whereas the WM representation of recently rewarded objects is noninformative because (1) it can only make a small contribution as the

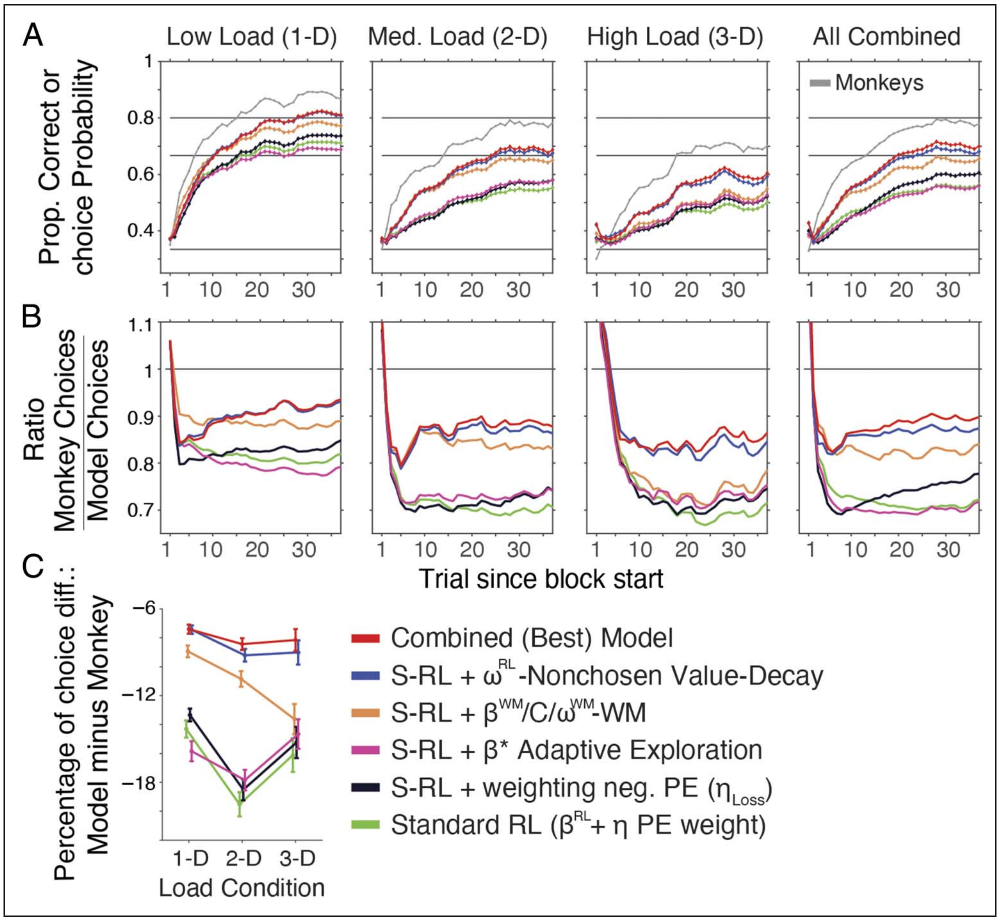


Figure 4. Choice probabilities of monkeys and models at three different loads. (A) Average choice accuracy of monkeys (gray) and choice probabilities of six models. The top-ranked model (red) combines WM with RL and selective suppression of nonchosen values, a separate learning gain for negative RPEs, and adaptive exploration rates. The base RL model (green) only contained a softmax beta parameter and a single learning rate. The other models each add a single mechanism to this base model to isolate its contribution to account for the choice patterns of the monkeys. Columns show from left to right the results for low-, medium-, and high-load conditions and for their average. (B) The ratio of monkey accuracy and model choice probability shows that, in all load conditions, the top-ranked model predicts monkey choices consistently better than models with a single added mechanism. (C) Average difference of model predictions (choice probability) and monkeys' choices (proportion correct) at low to high loads for different models. Error bars indicate *SE*. Prop. = proportion; diff. = difference.

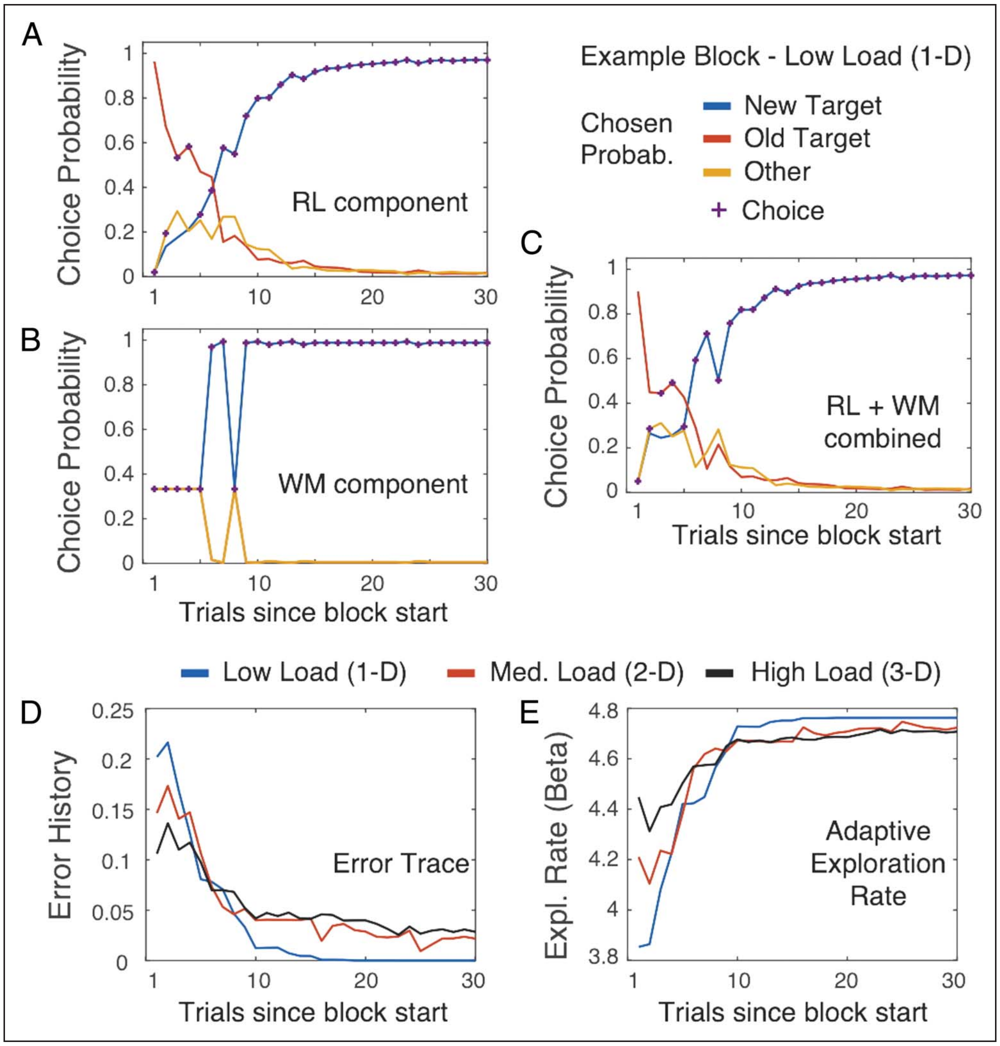


Figure 5. Contribution of WM, RL, and adaptive exploration to learning behavior. (A) Choice probabilities of the RL component of the top-ranked model for an example block, calculated for the objects with the new target feature (blue), the previous block's target feature (red), and other target features (yellow). Purple plus signs show which object was chosen. (B) Same format and same example block as in A but for choice probabilities calculated for objects within the WM module of the model. Choice probabilities of the WM and RL components are integrated to reach a final behavioral choice. (C) Same as A and B but after combining the WM and RL components in the full model. Choices closely follow the RL component, but when the WM representation is recently updated, its high choice probabilities influence the combined, final choice probability, as evident in Trials 6 and 7 in this example block. (D) The trace of nonrewarded (error) trials for three example blocks with low, medium, and high load peaks immediately after the block switch and then declines for all conditions. Error traces remain nonzero for the medium and high conditions. (E) The same example blocks as in D. The adaptive exploration rate (y axis) is higher (lower beta values) when the error trace is high during early trials in a block. Probab. = probability; Med. = medium.

number of stimuli in the block is much larger than the capacity and (2) it does not remember rewarded objects long enough to be around when the objects are presented another time.

Although the WM contribution declined with load, the ability to flexibly adjust exploration rates became more important with high load as is evidenced by improved choice probabilities at high loads (Figure 4C). This flexible meta-learning parameter used the trace of recent errors to increase exploration (reflected in lower beta parameter values). Such increases in exploration facilitate disengaging from previously relevant targets after the first errors after the block switch, even when there are no other competitive features in terms of value, because the mechanism enhances exploring objects with previously nonchosen features. Our results suggest that such an adjustment of exploration can reduce the decline in performance at high attentional loads (Figure 4C), that is, when subjects have to balance exploring the increased number of features with acting based on already gained target information (Figure 5D and E).

The Relative Contribution of Model Mechanisms for Learning (Exploration) and Plateau Performance (Exploitation)

The relative contributions of individual model mechanisms for different attentional loads can be inferred from their load-specific parameter values that best predicted monkeys' learning when fitted to the learning performance at each load separately (Figure 6). WM was maintained longer for learning at 2-D and 3-D than for 1-D load (lower ω^{WM} values for 2-D and 3-D, Figure 6C), but showed lower WM capacity relative to the number of active features (C_{WM}, Figure 6B) at 2-D and 3-D signifying that WM representations stopped contributing to learning at these loads. When attentional load increased, the models showed a gradual decrease of the weighting of positive PEs (η_{Gain} from ~0.2 to 0.1) and of the weighting of negative PEs (η_{Loss} from ~0.9 to 0.4; Figure 6E and G). A potential explanation for the decrease in η_{Gain} is that, with more distracting features, more trials are needed to determine what feature is the target, which can be achieved with slower updating. The decay of nonchosen feature values (ω^{RL}) was weaker with increased load across monkeys, indicating a longer retention of values of nonchosen objects (Figure 6F), which reflects protecting the target value when it is not part of the currently chosen (but unrewarded) object—an event that occurs more often at high loads. Adaptive exploration rates (β_m) increased on average from low to medium and high loads (more negative values) signifying increased exploration after errors at these higher attentional loads.

The parameter variations at increasing load could relate to either the learning speed or the plateau performance differences at different loads. To quantify their relative contributions, we used LME modeling to quantify how a model-independent estimate of learning speed (number of trials to reach criterion performance) and plateau accuracy (proportion of correct trials after learning criterion was reached) was predicted by the model parameters of the best-fit model. We found learning speed was significantly predicted by three parameters (Figure 7A). Learning occurred significantly earlier (i) with larger PE weighting for rewarded trials (η_{Gain}, t stat = −3.39, p = .0096, FDR controlled at alpha = .05), with higher PE weight for unrewarded trials (η_{Loss}, t stat = −4.66, p = .0016, FDR controlled at alpha = .05), and (iii) with larger adaptive change of exploration as captured in the meta-learning parameter β_m (t stat = −5.78, p = .00041, FDR controlled at alpha = .05; for scatterplot overviews, see Figure 7C). The remaining parameters were not significantly predicting learning when the FDR was controlled at an alpha of .05 (all not significant: β^{WM}: t = −2.11, C_{WM}: t = 0.33, ω^{WM}: t = −2.7, ω^{RL}: t = 0.41, α_+: t = 1.84).

In contrast to the learning speed, the plateau performance level was not significantly predicted by a single parameter when controlling for the FDR at an alpha of .05 (β^{WM}: t stat = 0.97, C_{WM}: t stat = −2.99, ω^{WM}: t = −0.11, β_m: t stat = −1.06, ω^{RL}: t = 2.32, η_{Gain}: t stat = −0.33, η_{Loss}: t stat = −1.34, α_+: t = 0.24). With a more lenient FDR of alpha = .2, the plateau performance was significantly predicted by parameter values of WM

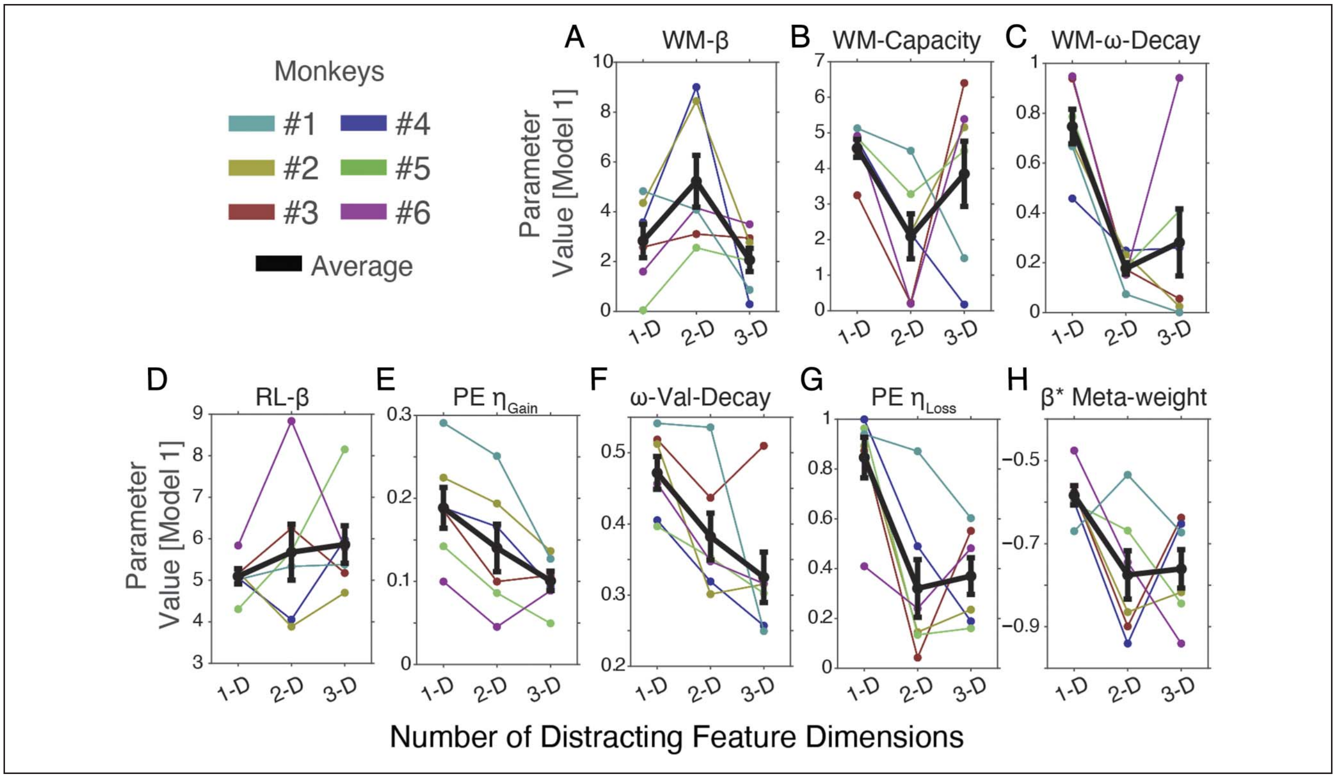


Figure 6. Model parameter values at different attentional loads. The average parameter values (black) of the top-ranked model (y axis) plotted against the number of distracting feature dimensions for the WM parameters (A–C) and the RL parameters (D–H). Individual monkeys are in colors. Error bars indicate *SE*.

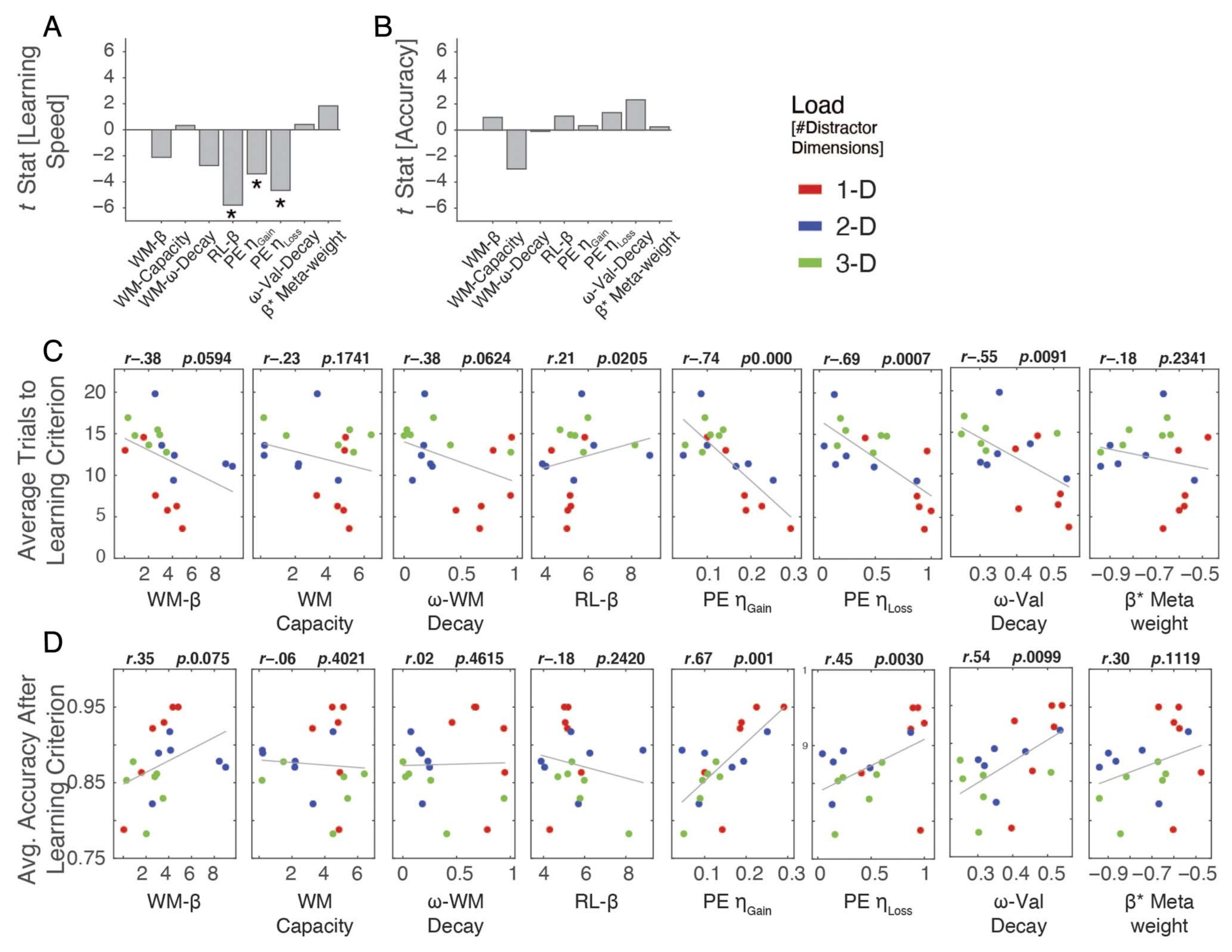


Figure 7. Model parameter values underlying learning speed and plateau performance levels. (A) The t values of the linear effects analysis for each parameter value of the best-fitting model for predicting the average trials-to-criterion (learning speed). Stars denote FDR-corrected significance at $p < .05$. Negative values denote that higher parameter values associate with faster learning. (B) Same format as A for the LME model predicting plateau performance accuracy with model parameter values. (C, D) Scatterplots of the RL parameter values of the top-ranked model plotted against the average learning speed (C) and the average plateau performance (D). The gray line is the linear regression whose r and p values are given above each plot. Each dot is the average result from one monkey in either the 1-D (red), 2-D (blue), or 3-D (green) condition. Val = value.

capacity (C_{WM}) and selective decay of nonchosen values (ω^{RL}; Figure 7B and D). These results indicate a trend for better performance with less reliance on WM, but with a stronger selective forgetting (decay) of values for features that were not part of chosen objects (modulated via ω^{RL}). Please note that Figure 7D and 7E also shows the correlation coefficients and the uncorrected p-value of the correlation that does not take into account the random effects variables that the linear mixed effects model results accounts for.

Model Parameter Values Distinguish Fast and Slow Learners

We next tested which model parameters distinguished good and bad learners across attentional load conditions by sorting subjects according to their learning speed, that is, their average number of trials to reach criterion, and predicting the rank order of fast to slow learners based on the parameter values of the best-fitting model (Figure 8A). We found that the variations of five model parameters had a significant main effect of the LME model (Figure 8B). Faster learners, requiring fewer trials to learn, retained a higher WM capacity (C_{WM}: $t = -3.12$, $p = .0124$; Figure 8C), a lower average exploration rate (β_m: $t = -5.33$, $p = .0005$; Figure 8D), a larger learning rate for positive outcomes (η_{Gain}: t stat $= -3.06$, $p = .0137$; Figure 8E), a larger learning rate for negative outcomes (η_{Loss}: t stat $= -4.08$, $p = .0028$; Figure 8F), and a larger variation (i.e., a larger amplitude of changes) of exploration rates (α_+: t stat $= -3.6$, $p = .0137$; Figure 8G). These findings illustrate that good and bad learners are

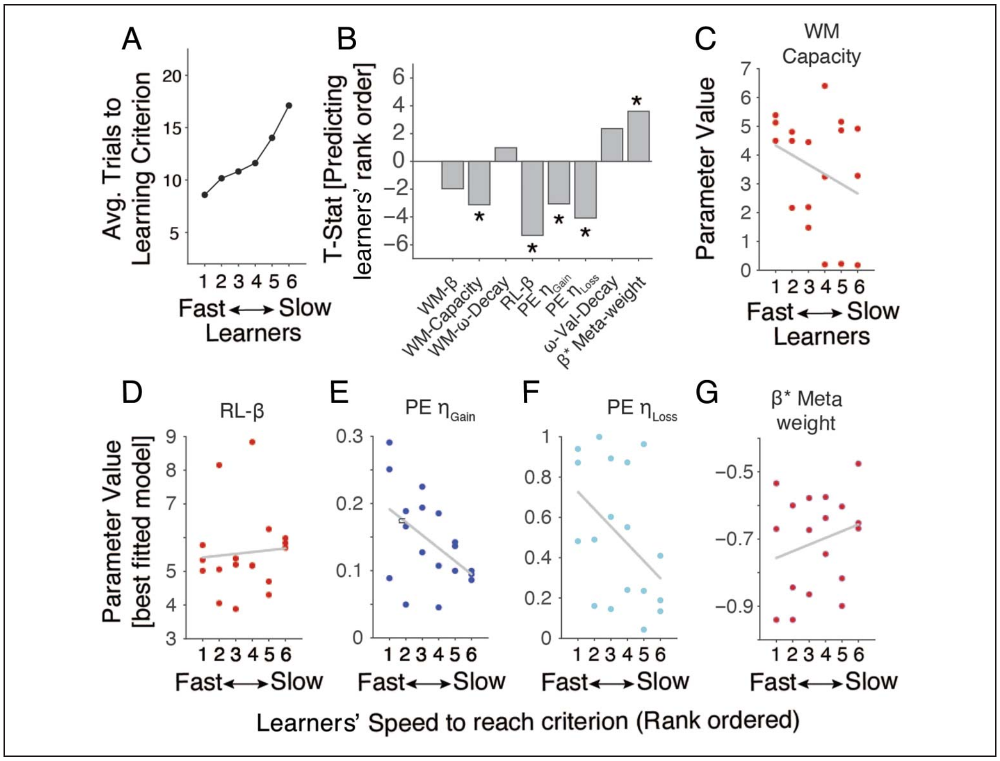


Figure 8. Model mechanisms distinguishing slow and fast learners. (A) The average learning speed (the trials to reach criterion; *y* axis) plotted against the individual monkeys ordered from the fastest to the slowest learner. (B) *t* Values of the linear effects analysis that tested how the rank-ordering of monkeys' learning speeds (Ranks 1–6) are accounted for by model parameter values. Stars denote FDR-corrected significance at $p < .05$. (C–G) The parameter values (*y* axis) of the best-fit model plotted against the rank ordering of learners (*x* axis). The parameters shown had a significant main effect to account for the rank ordering of learners as shown in B.

distinguished not by differences of a single learning mechanism but by applying a learning strategy that utilizes WM, flexibly adapts exploration rates, and shows enhanced learning rates for both correct and error outcomes (Figure 8).

DISCUSSION

We found that learning feature values under increasing attentional load are accounted for by an RL framework that incorporates four nonstandard RL mechanisms: (i) a value-decrementing mechanism that selectively reduces the feature values associated with the nonchosen object, (ii) a separate WM module that retains representations of rewarded objects over a few trials, (iii) separate gains for enhancing values after positive PEs and for suppressing values after negative PEs, and (iv) a meta-learning component that adjusts exploration levels according to an ongoing error trace. When these four mechanisms were combined, the learning behavior of monkeys was better accounted for than when using fewer or different sets of mechanisms. Critically, the same set of mechanisms was similarly important for all six animals (Figure 3), suggesting they constitute a canonical set of mechanisms underlying flexible learning and adjustment of behavior. Although subjects varied in how these mechanisms were weighted (Figure 6), those with faster learning and hence higher cognitive flexibility were distinguished by stronger weighting of positive and negative PEs, higher WM capacity, and an overall lower exploration rate but with enhanced meta-adjustment rates of the exploration rate during periods of high error rates. Taken together, these results document a formally defined set of mechanisms that support flexible learning of feature relevance under variable attentional load. It is important to note that the optimal model parameter settings do not perfectly account for all the observed choices; for instance, the observed learning curves in Figures 3C and 4A lie above those generated by the models. It is therefore possible that an additional model mechanism exists that closes this predictive gap, which could potentially interfere with, for instance, the interaction between WM- and RL-based learning during varying attentional load, hence potentially changing the suggested role of the mechanisms. Further research is therefore necessary to identify additional model mechanisms to exclude this possibility. In addition, these suggested mechanisms serve as a starting point for electrophysiological experiments in which specific brain areas are targeted by perturbative approaches to causally establish the role of model mechanisms in behavior. In the following, we further discuss our results from this viewpoint, including the evaluation of other model mechanisms we considered.

Selective Value Enhancement Is a Key Mechanism to Cope with High Attentional Load

One key finding was that only one nonstandard RL mechanism, the decay of values of nonchosen features

(ω^{RL}), contributed similarly to learning across all attentional load conditions (Figure 4C). This finding highlights the importance of this mechanism and supports previous studies that used a similar decay of nonchosen features to account for learning in multidimensional environments with deterministic or probabilistic reward schedules (Oemisch et al., 2019; Hassani et al., 2017; Radulescu et al., 2016; Niv et al., 2015; Wilson & Niv, 2011). The working principle of this mechanism is a push–pull effect on the expected values of encountered features and thus resembles a selective attention phenomenon (when emphasizing the "pushing" of values) of chosen and attended objects, or a "selective forgetting" phenomenon (when emphasizing the "pulling" down of values of nonchosen object features). When a feature is chosen (or attended), its value is updated and contributes to the next choice, whereas the value of a feature that is not chosen (not attended) is selectively suppressed and contributes less to the next choice. A process with a similar effect has been described in the associability literature whereby the exposure to stimuli without directed attention to it causes a reduction in effective salience of that stimulus. Such reduced effective salience reduces its associability and can cause the latent inhibition of nonchosen stimulus features for learning (Esber & Haselgrove, 2011; Donegan, 1981; Hall & Pearce, 1979) or the slowing of responses to those stimuli (also called negative priming; Lavie & Fox, 2000). The effect is consistent with a plasticity process that selectively tags synapses of those neuronal connections that represent chosen objects to enable their plasticity while preventing (or disabling) plasticity of nontagged synapses processing nonchosen objects (Roelfsema & Holtmaat, 2018; Rombouts et al., 2015). In computational models, such a synaptic tag is activated by feedback connections from motor circuits that carry information about what subjects looked at or manually chose (Rombouts et al., 2015). Accordingly, only chosen objects are updated, resembling how ω^{RL} implements increasing values for chosen objects when rewarded and the passive decay of values of nonchosen objects. At low attentional loads, high ω^{RL} values reflect the fast forgetting of nonchosen stimuli, whereas at high attentional loads, ω^{RL} adjusted to lower values, which is slowing down the forgetting of values associated with nonchosen objects (Figure 5F). The lowering of the ω^{RL} decay at high loads implies that values of all stimulus features are retained in the form of an implicit choice-history trace. Consistent with this finding, various studies have reported that several areas in prefrontal cortex contain neurons representing values of unchosen objects and unattended features of objects (Westendorff, Kaping, Everling, & Womelsdorf, 2016; Boorman, Behrens, Woolrich, & Rushworth, 2009). Our results demonstrate that, at high attentional loads, the ability of subjects to retain the value history of those nonchosen stimulus features is a critical factor for fast learning and good performance levels (Figure 7A).

WM Supports Learning Together with RL

Our study provides empirical evidence that learning the relevance of visual features leverages a fast WM mechanism in parallel with a slower RL of values. This finding empirically documents the existence of parallel (WM and RL) choice systems, each contributing to the monkey's choice in individual trials to optimize outcomes. The existence of such parallel choice and learning systems for learning fast and slow has a long history in the decision-making literature (Balleine, 2019; van der Meer, Kurth-Nelson, & Redish, 2012; Poldrack & Packard, 2003). For example, WM has been considered to be the key component for a rule-based learning system that uses a memory of recent rewards to decide to stay with or switch response strategies (Worthy, Otto, & Maddox, 2012). A separate learning system is associative and implicitly integrates experiences over longer periods (Poldrack & Packard, 2003), which in our model corresponds to the RL module.

The WM mechanisms we adopted for the feature learning task are similar to WM mechanisms that contributed in previous studies to the learning of strategies of a matching pennies game (Seo, Cai, Donahue, & Lee, 2014), the learning of hierarchical task structures (Alexander & Brown, 2015; Collins & Frank, 2012, 2013), or the flexible learning of reward locations (Viejo et al., 2018; Viejo et al., 2015). Our study adds to these prior studies by documenting that the benefit of WM is restricted to tasks with low and medium attentional loads (Figure 4). The failure of WM to contribute to learning at higher loads likely reflects an inherent limit in WM capacity. Beyond an interpretation that WM capacity limits are reached at higher loads, WM is functionally predominantly used to facilitate processing of actively processed items as opposed to inhibiting the processing of items stored in WM (Noonan, Crittenden, Jensen, & Stokes, 2018). In other words, a useful WM is rarely filled with distracting, nonrelevant information that a subject avoids. In our task, high distractor load would thus overwhelm the WM store with information about nonrewarded objects whose active use would not lead to reward. Consequently, the model—and the subject whose choices the model predicts—downregulated the importance of WM at high attentional loads, relying instead on a slower RL mechanism to cope with the task.

Separate Learning Rates Promote Avoiding Choosing Objects Resulting in Worse-Than-Expected Outcomes

We found that separating learning from positive and negative PEs improved model predictions of learning across attentional loads (Figure 4) by using considerably larger learning rates for negative than positive outcomes (Figure 6E vs. 6G). Thus, monkeys were biased to learn faster to avoid objects with worse-than-expected feature values than to stay with choosing objects with better-than-expected feature values. A related finding is the observation of larger learning rates for losses than gains for monkeys performing a simpler object–reward association task (Taswell et al., 2018). In our task, such a stronger weighting of erroneous outcomes seems particularly adaptive because the trial outcomes were deterministic, rather than probabilistic, and thus a lack of reward provided certain information that the chosen features were part of the distracting feature set. Experiencing an omission of reward can therefore immediately inform subjects that feature values of the chosen object should be suppressed as much as possible to avoid choosing it again. This interpretation is consistent with recent computational insights that the main effect of having separate learning rates for positive and negative outcomes is to maximize the contrast between available values for optimized future choices (Cazé & van der Meer, 2013). According to this rationale, larger learning rates for negative outcomes in our task promote the switching away from choosing objects with similar features again in the future. We should note that, in studies with uncertain (probabilistic) reward associations that cause low reward rates, the overweighting of negative outcomes would be nonadaptive as it would promote frequent switching of choices, which is suboptimal in these probabilistic environments (Cazé & van der Meer, 2013). These considerations can also explain why multiple prior studies with probabilistic reward schedules report an overweighting of positive over negative PEs, which in their tasks promoted staying with and prevent switching from recent choices (Lefebvre et al., 2017; van den Bos et al., 2012; Kahnt et al., 2009; Frank et al., 2004, 2007).

The separation of two learning rates also demonstrates that our task involves two distinct learning systems for updating values after experiencing nonrewarded and rewarded choice outcomes. Neurobiologically, this finding is consistent with studies of lesioned macaques reporting that learning from aversive outcomes is more rapid than learning from positive outcomes and that this rapid learning is realized by fast learning rates in the amygdala as opposed to slower learning rates for better-than-expected outcomes that are closely associated with the ventral striatum (Taswell et al., 2018; Averbeck, 2017; Namburi et al., 2015). Our finding of considerably higher (faster) learning rates for negative than positive PEs is consistent with this view of a fast versus slow RL updating system in the amygdala and the ventral striatum, respectively. The importance of these learning systems for cognitive flexibility is evident by acknowledging that learning rates from both, positive and negative outcomes, distinguished good and bad learners (Figure 8), which supports reports that better and worse learning human subjects differ prominently in their strength of PE updating signals (Krugel, Biele, Mohr, Li, & Heekeren, 2009; Klein et al., 2007; Schönberg, Daw, Joel, & O'Doherty, 2007).

Adaptive Exploration Contributes to Learning at High Attentional Load

We found that adaptive increases of exploration during the learning period contributed to improved learning at high loads (Figure 3). Adapting the rate of exploration over exploitation reflects a meta-learning strategy that changes the learning process itself by adaptively enhancing searching for new choice options irrespective of already-acquired expected values (Doya, 2002). Our finding critically extends insights that adaptive learning rates are critically important to cope with uncertain environments (Soltani & Izquierdo, 2019; Farashahi, Donahue, et al., 2017) to target uncertainty imposed by increased distractor load. In earlier studies, reward uncertainty was estimated to adjust learning rates in tasks with varying volatility (Farashahi, Donahue, et al., 2017), changing outcome probabilities when predicting sequences of numbers (Nassar, Wilson, Heasly, & Gold, 2010), sharp transitions of exploratory search for reward rules and exploitation of those rules (Khamassi et al., 2015), probabilistic reward schedules during reversal learning (Krugel et al., 2009), or the compensation for error in multijoint motor learning (Schweighofer & Arbib, 1998). A commonality of these prior meta-learning studies is a relatively high level of uncertainty about the source of reward or error outcomes. In our task, the uncertainty about the target feature systematically increased with the number of distracting features. As a consequence of enhanced uncertainty, subjects utilized a learning mechanism that increased randomly exploring new choice options when nonrewarded choices accumulated and to reduce exploring alternative choices when choices began to lead to reward outcomes. Such balancing of exploration and exploitation can be achieved by using a memory of recent reward history to adjust undirected vigilance (Khamassi et al., 2013; Dehaene, Kerszberg, & Changeux, 1998) or other forms of exploratory strategies (Tomov et al., 2020).

Limitations and Scope of Our Findings

Our study found that four nonstandard learning mechanisms contributed to explaining learning at low and high attentional loads across six monkeys. This is a major novel finding that motivates identifying the neural basis of each of these mechanisms and how they cooperate during learning. However, these results should not be considered a conclusive list of learning mechanisms, and various limitations of our approach should be considered. First, we cannot rule out that other mechanisms are used beyond those considered (see next subsection). Second, our conclusions are based on ranking different models according to how well (in terms of likelihood) they predict choices. We quantified this for each of six monkeys separately with the BIC, which is well established and penalizes the model predictability by the number of parameters used for the prediction (Wagenmakers & Farrell, 2004). We could have chosen other means to rank order models by, for example, combining the BIC with measures of explained variance of choices (pseudo R^2) into a compound goodness-of-fit score (Balcarras et al., 2016) or performing a model recovery analysis using all possible models (Wilson & Collins, 2019). Such more extensive model comparisons will be justified when a study considers multiple alternative, mutually exclusive, and possibly more contentious, newly devised learning mechanisms that do not apparently contribute to enhancing likelihood in terms of BIC. Here, we documented, as reported in the Methods section, that the models used were identifiable to a sufficient degree so that the different parameters could be compared and that the models yielded significant differences in the objective function so that the performance of the models could be adequately compared (Appendix).

Consideration of Alternative Mechanisms

In our study, only two main mechanisms were tested that were not consistently contributing to enhancing the BIC (choice stickiness and Bayesian dimension weighting). These two processes might play a role in other tasks that we did not consider. Our study should therefore not be considered to exclude learning mechanisms but rather to provide strong positive empirical evidence for including those four mechanisms that we found to consistently contribute to successful cognitive flexibility in our task.

Beyond the mechanisms that we tested explicitly in our study, we explored alternative models that have been described in the literature and that we found not to be competitive for our data set. These models were almost exclusively formulated in terms of values, either of the object or feature. One of the omitted models was object-based RL learning (not to be confused with object-based WM, which represented fast, one-shot learning). For our data, this model did not yield competitive accounts for the choices. It is important to mention this model because it relates to the investigation in Farashahi, Rowe, et al. (2017), which compares feature with object-based probabilistic RL learning.

A second set of models we did not explicitly consider in this report but tested on the data set are models that assume learning involves subjects to test hypotheses of reward rules causing fast and discrete switches of attention and behavioral choices when a hypothesis is refuted. For example, some studies report that the identification of the correct target occurs suddenly during the block and from then on results in choices that are rewarded (Papachristos & Gallistel, 2006). From block to block, this sudden onset occurs at different trials relative to the target switch so averaging accuracy across blocks can give the wrong impression of a smooth learning curve. The rapid switching is consistent with the subject holding a single hypothesis at a time about what the target feature is and switching when the choices based on this hypothesis are not rewarded. Such a model is referred to as a serial hypothesis-testing model as suggested by Niv and colleagues (Radulescu, 2020; Radulescu, Niv, & Ballard, 2019). We implemented a version of this model, explored the learning behavior in the generative version of the model, and fitted it to a subset of behavioral data using the likelihood formulation of the model (see Appendix, Appendix Figure A2). On each trial, there is a single inferred target feature, and a switch to a new (randomly chosen) feature is made when the number of rewards in the prior τ trials is below a certain threshold (this is a parameter in the model). At the phenomenological level, the model reproduces the decrease in learning speed with higher attentional load (Appendix Figure A2D–F and Appendix Figure A3B–D for Attentional Loads 1, 2 and 3, respectively). For our fits, we varied the memory duration τ and found that a memory of three trials in the past was optimal; nevertheless, the resulting performance (NLL) did not match our best model (Model 1; Table 1). We expect that the model will be interesting for future studies in which probabilistic rewards are considered for which memory should help in distinguishing between unrewarded correct choices and unrewarded incorrect choices for switching to a new hypothesis.

A more complicated version of the hypothesis switching approach is based on change-point detection (Adams & MacKay, 2007), in the sense that more probability distributions need to be updated from trial to trial (hence be represented somewhere in the brain if the model has to have a mechanistic interpretation). A Bayesian inference model with change-point detection was described in Wilson and Niv (2011, Section 2.3.2). In this model, the key quantity is the probability p_f that feature f is the target. This distribution converges quickly toward a situation in which all the probability weight is on one feature, in essence representing that the subject has a single hypothesis for the target feature. A change point indicates that the current hypothesis does not predict rewarded choices anymore, in which case p_f has to be reset, typically to a uniform distribution. The model builds, based on the stimulus–choice–reward sequence, a probability distribution of the possible change points and maintains the corresponding feature distribution p_f conditioned on each possible change point in the past. We had earlier simulated a Bayesian inference model without change-point detection, which led to noncompetitive values for NLL, because it learned too fast compared to the subject (note that we had a deterministic reward rule). We had included this model, for example, in Oemisch et al. (2019). Similarly, the model with change-point detection also had a higher NLL than our best model, and the optimal fit parameters corresponded to an unreasonably high switch rate h. Hence, we did not enter these types of models in our comparison.

An alternative to value-based learning is Q-learning in which for each state s of the environment and inferred feature f, an action value $Q(c, f, s)$ for choice (action) c is learned from the past choices and rewards. An example of such a model, relevant to our data, was proposed in Kour and Morris (2019). An additional advantage was that this model was fit by an expectation–maximization procedure (i.e., Baum–Welch method; see Murphy, 2012) because that procedure provides a way to reconstruct the hypothesis on each trial. We however found that the sheer number of parameters represented by the Q values (which grow with the product of the number of states times the number of features times the number of objects, which in our case is 36 × 3 × 3, 216 × 6 × 3, and 1296 × 9 × 3 for one, two, and three nonneutral features, respectively) made fitting difficult for the number of blocks that we had experimentally available.

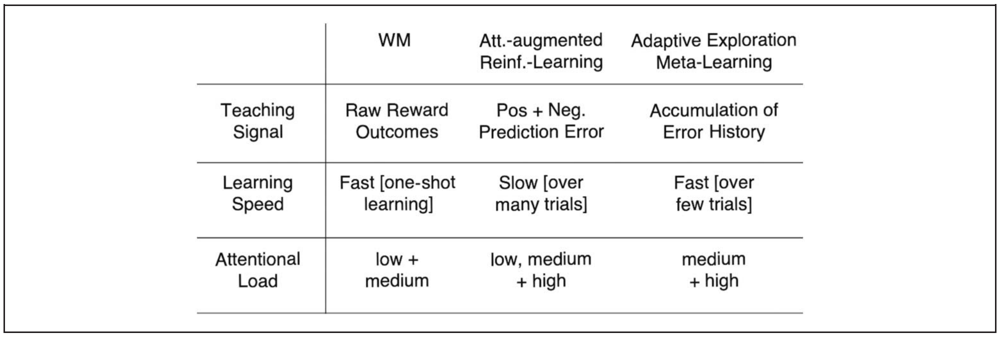
In other published models, the specific target of updates for unchosen features is a free parameter (Ito & Doya, 2009; Barraclough, Conroy, & Lee, 2004). These models are based on Q-learning hence suffer in our case from the aforementioned issue of a large number of parameters necessary for the action values. This was not the case in the cited references because there was only one state in combination with two possible choices (L and R) that each has a different reward probability associated with it. A direct comparison based on our data was therefore not feasible. Nevertheless, Ito and Doya (2009) discuss four different update rules, which determine how the action values of chosen and nonchosen options are updated when there was a reward or when the reward was omitted. These update choices play similar roles as our learning rate parameters η_{Loss} and η_{Gain} play for the update of the chosen option when not rewarded/rewarded, respectively, and ω_{nc}^{RL} for the decay of the nonchosen option. In conclusion, our evaluated models are similar in spirit to the ones in Ito and Doya (2009); however, rather than based on action values for choices, we update feature values with similar alternative update models.

Other published models incorporated longer-timescale perseveration to account for choice behavior (Miller, Shenhav, & Ludvig, 2019; Akaishi, Umeda, Nagase, & Sakai, 2014). For example, the task in Akaishi et al. (2014) involved two choices, but models for this choice were formulated solely in terms of the probability of making the first choice, which is obtained by transforming a decision variable by a sigmoid. Different choices for functional dependence of the decision variable on past choices and current and past stimulus features were made. For instance, it could depend on the prior choice, the current and previous contrast, sometimes in a nonlinear combination, where the previous contrast gain modulated the effect of the current contrast. We have incorporated the choice stickiness of the previous choice in a similar fashion (following Balcarras et al., 2016) but found that, for the current task, stickiness on average did not improve the prediction of choices and was not part of the top-ranked model in five of six monkeys (Figure 3A). Akaishi and colleagues also evaluated replacing variables by predictions for the stimulus contrast and choice (Akaishi et al., 2014). These predictions were updated using the PE similar to the feature values in our RL models, hence this mimics Q-learning. Apart from the fact we did not have a contrast variable in the current behavioral data, we did not include such prediction variables in our models, primarily because choices depend on the stimulus configuration, hence we would have to consider a different set of action variables for each stimulus configuration, again yielding the aforementioned state-space size problems.

Taken together, our brief discussion of previous work covers models that either we implemented in a pilot stage but found not to account well for the choices in our task or were a different type of model that was based on action value formulation that was feasible in the original works, but not for our behavioral data. The latter incorporated effects of perseveration and different reward-dependent gains and choice-dependent decay that we had already incorporated in our feature value updates and evaluated as part of our set of models. Taken together, we believe that our comparison covered a sufficiently diverse set of models to draw conclusions about how behavioral strategies change with attentional load. When the task is extended to include probabilistic rewards as well as multistage setups, then action-value-based models may need to be used instead of the feature-value-based models we focused on here.

Conclusion

In summary, our study documents that a standard RL modeling approach does not capture the cognitive processes needed to solve feature-based learning. By formalizing the subcomponent processes needed to augment standard (Rescorla–Wagner) RL modeling, we provide strong empirical evidence for the recently proposed "EF-RL" framework that describes how executive functions (EF) augment RL mechanism during cognitive tasks (Rmus, McDougle, & Collins, 2020). The framework asserts that RL mechanisms are central for learning a policy to address task challenges but that attention-, action-, and higher-order expectations are integral for shaping these policies (Rmus et al., 2020). In our study, these "EF" functions included (i) WM, (ii) adaptive exploration, (iii) a separate learning gain for erroneous performance, and (iv) an attentional mechanism for forgetting nonchosen values. As outlined in Figure 9, these mechanisms leverage distinct learning signals, updating values based directly on outcomes (WM), on PEs (RL-based decay of nonchosen values), or on a continuous error history trace (meta-learning-based adaptive exploration). As a



	WM	Att.-augmented Reinf.-Learning	Adaptive Exploration Meta-Learning
Teaching Signal	Raw Reward Outcomes	Pos + Neg. Prediction Error	Accumulation of Error History
Learning Speed	Fast [one-shot learning]	Slow [over many trials]	Fast [over few trials]
Attentional Load	low + medium	low, medium + high	medium + high

Figure 9. Characteristics of the WM, RL, and meta-learning components. The model components differ in the teaching signals that trigger adjustment (top row), in the learning speed, that is, in how fast they affect behavior (center row) and in how important they are to contribute to learning at increasing attentional load (bottom row). Att. = attention; Reinf. = reinforcement; Pos = positive.

consequence, these three learning mechanisms operate in parallel and influence choices to variable degrees across different load conditions, for instance, learning fast versus slow (WM vs. meta-learning vs. RL) and adapting optimally to low versus high attentional load (WM vs. meta-learning). Our study documents that these mechanisms operate in parallel when monkeys learn the relevance of features of multidimensional objects, providing a starting point to identify how diverse neural systems integrate these mechanisms during cognitive flexible behavior (Womelsdorf & Everling, 2015).

APPENDIX

Identifiability of Model 1

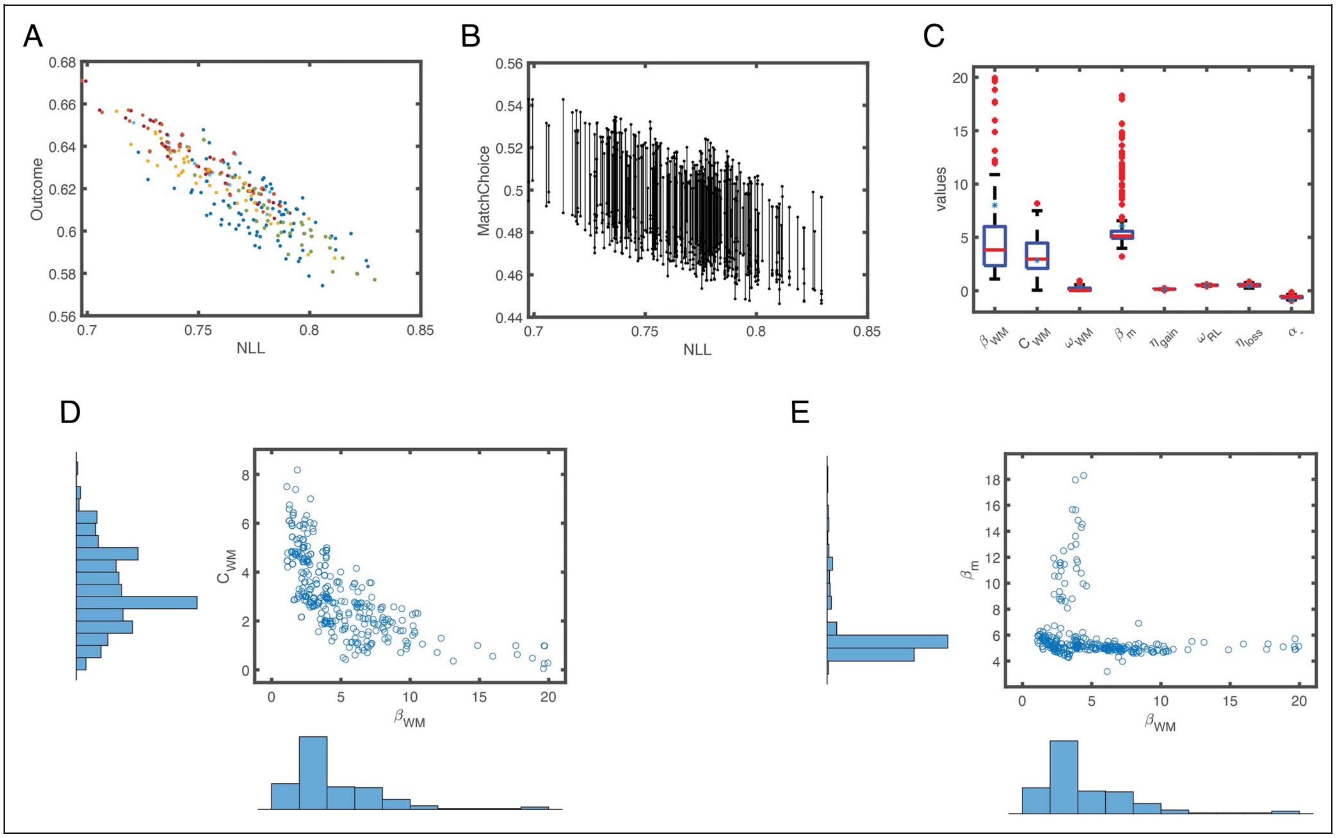
We evaluated to what extent the fitting of Model 1 (Table 1) was affected by local minima by starting the fitting procedure from different initial conditions (n = 10). We found that a number of distinct parameter sets were reached multiple times. The objective function, NLL, was different for these solutions, but the differences were small and typically less than the difference between different models considered here. For each of the initial conditions, we generated 40 behavioral sequences and determined the NLL, reward outcome, and match between subject and model choices (Appendix Figure A1A and B). Here, the match was quantified as the fraction of common choices. The overall conclusion is that the better the NLL (i.e., a lower value), the more overlap there was between the choices of the model and those of the subject; this was, in addition, indexed as a higher average reward. We also refitted these sequences and analyzed the variability in parameter values so obtained, represented as boxplots in Appendix Figure A1C. We found that the variability in the β parameters for both the WM and RL components was high, which was reflected in the negative correlation between them (Appendix Figure A1E). In addition, there was a strong anticorrelation between β_{WM} and C_{WM} (Appendix Figure A1D). This suggests that outlier values of the β variables should be removed. This can be achieved by putting priors on the parameters and incorporating those in the objective function or using a cross-validation strategy to extract the robust parameter sets.

Serial Hypothesis-Testing Models

Generative Model

The state of the subject is described by a latent variable z; in our case, this is the current hypothesis about what the target feature f is. There are multiple dimensions d; within each dimension, there are multiple feature values f_d, the number of which varies with dimension. For instance, for the dimension “color,” there are feature values “red” and “green,” whereas for the dimension “shape,” there are a different number of features. Each dimension has a neutral feature value that can appear in multiple objects, when the corresponding dimension cannot provide the target feature that predicts reward. The key variable is the target feature f_t during trial t, determined by the environment, and the set of features that are active, f_a; that is, they are all the nonneutral features that are presented during a block of trials and that can be the target feature. We use functions to switch between dimension d (color), feature value within dimension f_d (red), and overall feature value f (index for color–red), specifically $D(f) = d_f$, $F_d(f) = f_d, f = F(d, f_d)$; these are easily implemented as a lookup in the appropriate matrix. The primary reason is that, unless you use attention to a particular dimension, the overall feature value can be used as the main variable in updates, and then, it also represents the values z can take.

The state is an allowed combination of n_{o} objects, and the objects can share the same feature value for a dimension when it is the neutral value, but they have to have different ones for nonneutral features. The state of the environment is summarized by the matrix $S(i, j, s)$, here $i = 1, \ldots, n_{\mathrm{o}}$ indicates the object, $j = 1, \ldots, n_f$ indicates the feature, and $s = 1, \ldots, n_s$ indicates the state index. When in trial t the state of the environment is s_t, it means that the objects are given in terms of their features by $S(i, j, s_t)$, that is, $S(i, j, s_t) = 1$ when object i in state s_t contains feature j. Taken together, this means that the environment can be completely described by state s_t (n_s = 36, 216, or 1296, respectively, for one, two, or three active dimensions with three nonneutral features used in each dimension) and the current target feature f_t (taking values between 1 and n_f = 3, 6 and 9, respectively, for the aforementioned attentional loads). In our setup, each s_t is chosen randomly from among the n_s available states,



Appendix Figure A1. Model 1 identifiability. The experimental behavioral data were fitted 10 times starting from different initial conditions for the parameters. For each of the fit parameters, the generative model was run 40 times, which yielded 400 values for the NLL, the average reward, and the match between subject choices. The models were refitted, which yielded 400 parameter sets. (A) Reward outcome versus NLL: A lower (better) NLL also gives a higher average reward. (B) Match between model and subject choices versus NLL. Each model choice sequence was represented by a line; the bottom point is the match for randomized model choices, whereas the top point is that for the actual model choices. The lower the NLL, the better the prediction of choices by the model is. (C) Boxplot summarizing the 400 values of each parameter; the red line indicates the median, the box contains the interquartile range (25th–75th percentile), and the points beyond the whiskers are outliers. For β_{WM} and β_m, there are a significant number of outliers. The variables (D) β_{WM} and C_{WM} and (E) β_{WM} and β_m are negatively correlated.

where f_t switches randomly from trial to trial from feature j to i, according to the switching matrix

$$p_{ij}^{swf} = (1-h)\delta_{ij} + \frac{h}{n_f - 1}\left(1-\delta_{ij}\right),$$

where δ_{ij} is the Kronecker delta, equal to 1 when $i = j$ and zero otherwise. Hence, there is a probability h for a switch to a random other feature. The potential value (possible reward) of each state is the same, hence the subject cannot change future values of a state by a particular choice. This makes this system different from the typical one studied using Q-learning.

The additional variables on each trial are subject-inferred target z_t, choice c_t, and the resulting reward r_t. The choice c_t is for the object that contains the inferred target, hence, for which $S(c_t, z_t, s_t) = 1$, with probability p_c (exploitation) and with $1 - p_c$, a random choice is made (exploration). This parameter, in some sense, plays the role of the β in the softmax choice probability of our feature value models (Table 1). The probability for obtaining a reward on trial t, $r_t = 1$, is p_r when the choice contains the target, that is, $S(c_t, f_t, s_t) = 1$, and p_{nr}, otherwise.

In the experimental study, we used a deterministic reward scheme, $p_r = 1$ and $p_{nr} = 0$. Taken together, using $x = S(c_t, f_t, s_t)$ as shorthand, then the probability for reward on trial t is $p = p_r x + p_{nr}(1 - x)$. Note that c_t and r_t can only influence the future inferred target z_{t+1}.

The probability for the new inferred target being $z_{t+1} = i$ when $z_t = j$ is determined by the following switching matrix

$$p_{ij}^{swz} = \left(1-h_t^z\right)\delta_{ij} + \frac{h_t^z}{n_f - 1}\left(1-\delta_{ij}\right),$$

with the switch probability

$$h_t^z = 1 - \frac{1}{1+e^{-\kappa(\bar{r}_t - \theta)}}.$$

Note that the sigmoid in this expression in fact represents a stay probability, but we need to enter the switch probability in the switch matrix, hence the subtraction from one in the formula. The $\bar{r}_t$ is the weighted average of rewards across the past τ trials, $\bar{r}_t = \sum_{i=0}^{\tau-1} w_i r_{t-i}$; here, we explored two choices: $w_i = 2^{\tau-i-1}$ or $w_i = 1$. Hence, either the more recent reward counts more or all past rewards are counted equally.

The algorithm to create the behavioral model data therefore has the following steps:

- Initialize: Set z_1, f_1, and s_1 (drawn uniformly from among all feasible values).
- Make choice c_t: Choose whether to exploit (probability p_c) or explore $(1 - p_c)$; for the former, choose the c_t that satisfies $S(c_t, z_t, s_t) = 1$, whereas for the latter, choose c_t according to the uniform distribution.
- Determine reward r_t: Set $r_t = 1$ with probability $p = p_r x + p_{nr}(1 - x)$, $x = S(c_t, f_t, s_t)$.
- Update inferred target z_t: Draw $z_{t+1} = i$ according to p^{swz}_{i,z_t} using

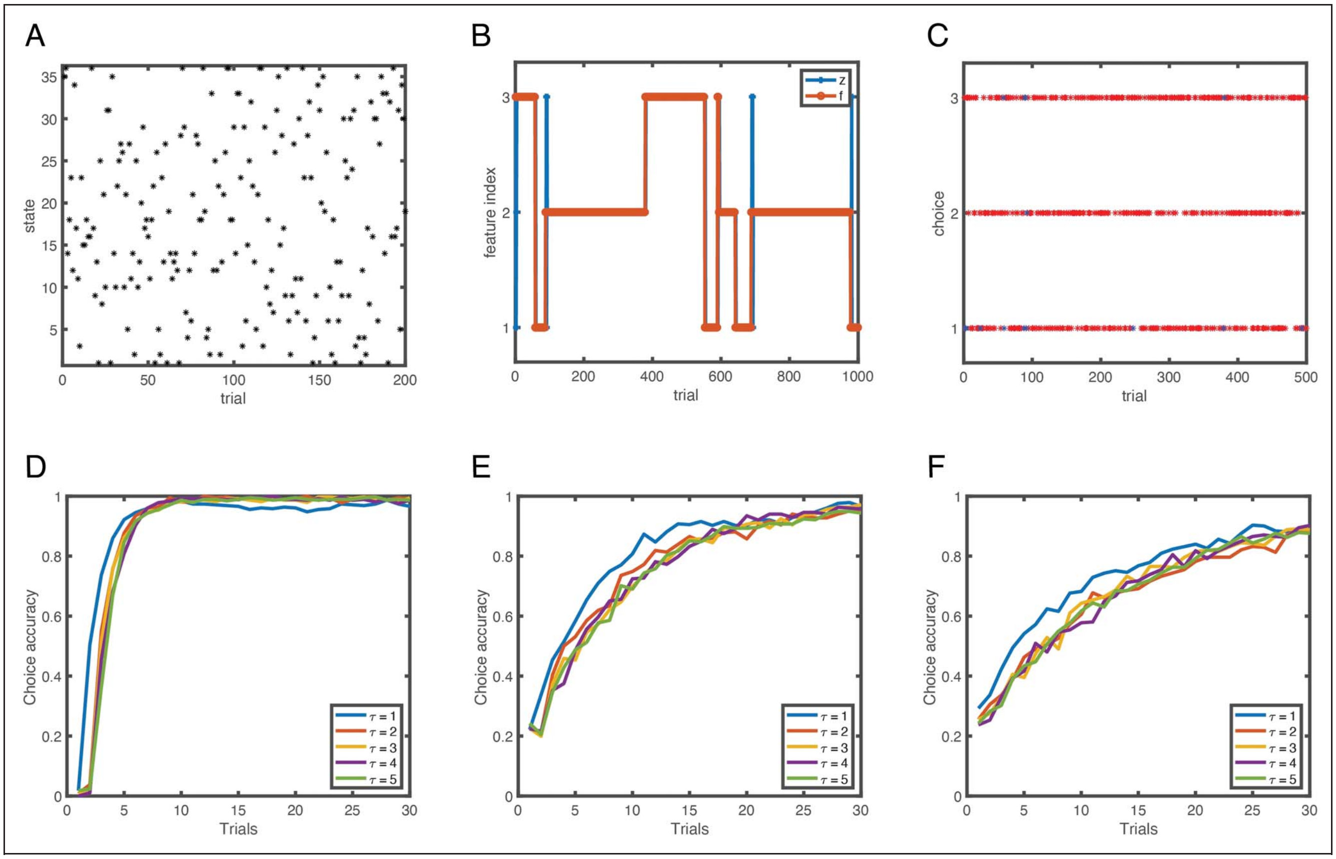
$$h_t^z = 1 - \frac{1}{1 - e^{-\kappa(\bar{r}_t - \theta)}}.$$

- Update target feature f_t: Draw $f_{t+1} = i$ according to p^{swf}_{i,f_t}.
- Update state (objects) s_t: Draw s_{t+1} uniformly from among the allowed states.

The parameters are exploit probability p_c, reward probability for correct choice p_r and for incorrect choice p_{nr}, target switching rate h, memory duration τ, sharpness of switch function κ, and the threshold θ.

Likelihood Model for Behavioral Observations

The generative model produces a sequence (s_t, c_t, r_t), and the actual target f_t is also available but should not be used (because this is contained in the reward that the subject receives), whereas latent variable z_t is hidden. During the fitting procedure, we fixed p_r and p_{nr}, although in principle, the subject may not know them, and we assumed a deterministic choice $p_c = 1$. In addition, the allowed states according to which the sequence was generated are also given; this means a fixed attentional load is assumed and provided. Hence, the free



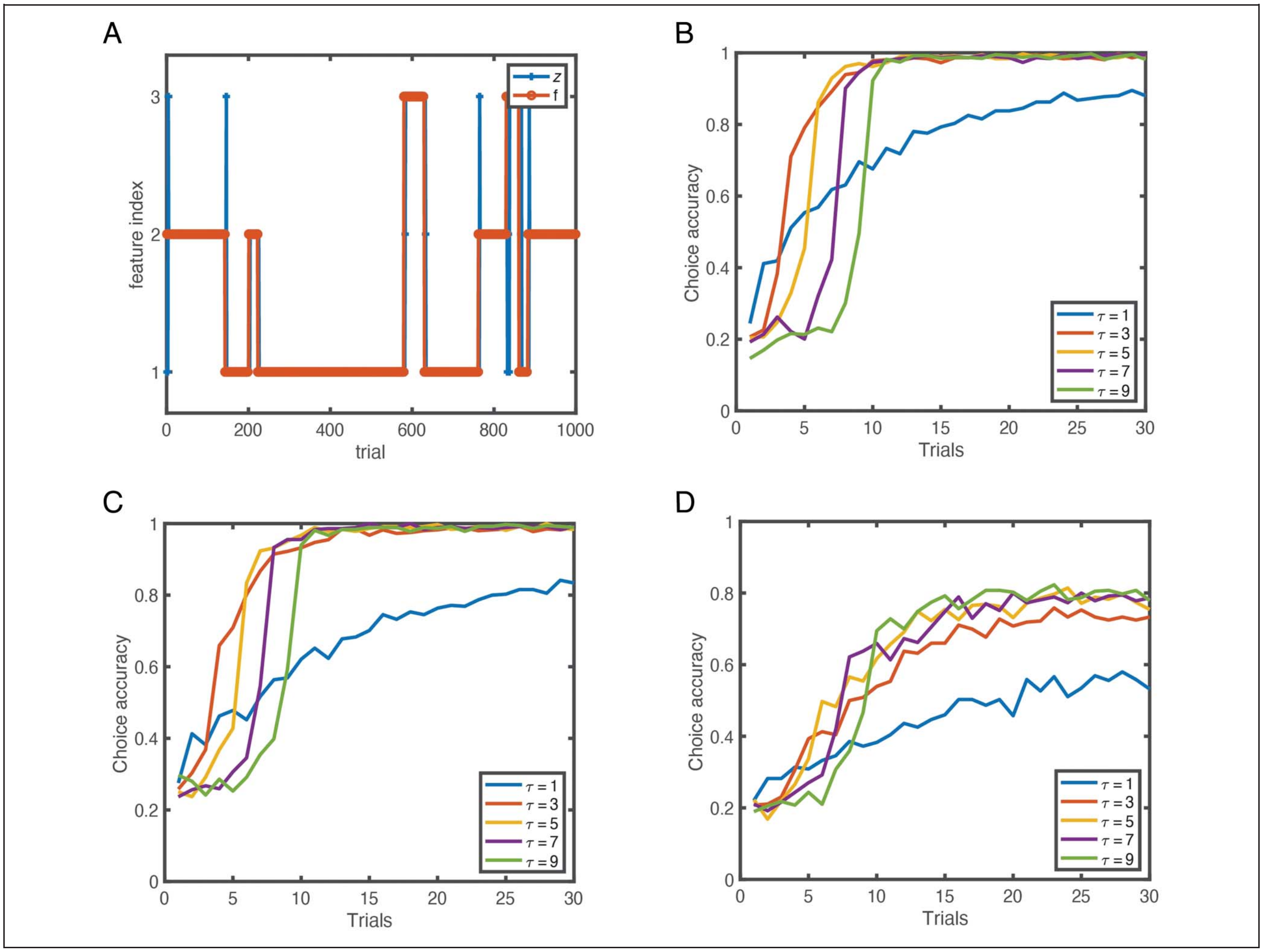
Appendix Figure A2. Behavioral data for a serial hypotheses-testing model with $p_r = 0.99$, $p_{nr} = 0$, $p_c = 1$, and $h = 0.01$, with recent rewards more heavily weighted in the switching function. (A) There are $n_s = 36$ different states, which are randomly sampled across trials. (B) The target feature f_t is generated from a random switching process with hazard rate $h = 0.01$; the target feature z_t is inferred by the model based on previous observations of choices and reward. (C) The choice as a function of trial index: Rewarded choices are in red; and the unrewarded, in blue. These examples are generated for $\tau = 3$, $\kappa = 5.71$, $\theta = 1.75$, and Attentional Load 1. (D–F) Choice accuracy curves for attentional load equal to (D) 1, (E) 2, and (F) 3, for five different values of $\tau = 1, \ldots, 5$ as indicated in the legend. The corresponding (κ, θ) values are (40, 0.25), (13.33, 0.75), (5.71, 1.75), (2.67, 3.75), and (1.3, 7.75). Higher attentional load leads to slower learning and lower asymptotic choice accuracy. For these settings, there is little advantage of a longer memory (τ); for Attentional Load 1, longer τ prolongs the learning period.

parameters are κ and θ, which can be optimized for a relevant range of memory durations τ. We start the procedure by setting the inferred target to a specific value z_t. Then, on the basis of the reward history, a switch probability h_t^z and the resulting switch vector p_{i,z_t} (given by p^{swz}_{i,z_t}, i runs along allowed feature values) are formed. The vector $S(c_{t+1}, i, s_{t+1})$ indicates which features were present in the chosen object on the next trial, hence the product of the switching probability to i and whether the choice contained i measure how well the new inferred target predicts the choice. We choose as log likelihood $LL = log(\Sigma_i\, S(c_{t+1}, i, s_{t+1})p_{i,z_t})$ and add these across trials. We also need to update the inferred target; we choose the $z_{t+1} = i$ where i maximizes $S(c_{t+1}, i, s_{t+1})p_{i,z_t}$. This procedure recovers the correct parameters when fed the model-generated choice–state–reward sequence.

To fit the experimental data, which contain blocks with varying attentional loads, we add a running average across presented features to construct the correct set of active feature values f_a for the switching functions.

Serial Hypothesis Testing Based Update with Feature-specific Switching

The preceding model switched when there were too many unrewarded trials in its past, but it did not use that information to switch to a specific feature. In a different version of the model, the inferred target τ trials back was used as the initial condition for a Bayesian update that integrated the choice–state–reward sequence up to and including the current trial. Specifically, set $p_i^f = \delta_{i,z_{t-\tau}} + 0.01$, for $i = 1, \ldots, n_f$, as starting point of the iteration. We add a small nonzero probability for all other features, because otherwise, with the multiplicative updates used here, there can never appear any nonzero probabilities for other features than the starting one. The update for



Appendix Figure A3. Behavioral data for a serial hypotheses-testing model with feature-based switching, $p_r = 0.99$, $p_{nr} = 0$, $p_c = 1$, and $h = 0.01$. (A) The target feature f_t is generated from a random switching process with hazard rate $h = 0.01$; the target feature z_t is inferred by the model based on previous observations of choices and reward. These data are for τ = 3 and Attentional Load 1. (B–D) Choice accuracy curves when the attentional load is (B) 1, (C) 2, and (D) 3, for five different values of τ = 1, 2, …, 9 as indicated in the legend. Higher attentional load leads to slower learning and lower asymptotic choice accuracy. Longer τ prolongs the learning period.

each trial is, using temporary variables x, y, and u for ease of presentation, and v as temporary trial index: $x_i = \delta_{i,c_v}$ and $y_{ij} = S(i, j, s_v)$, for $i = 1, \ldots, n_o$, and $j = 1, \ldots, n_f$, whereas $u_j = \Sigma_i[p_r x_i y_{ij} + p_{nr}(1 - x_i)(1 - y_{ij})]$. The update then becomes $p_j^f = [r_v u_j + (1 - r_v)(1 - u_j)]p_j^f$, which is applied starting from the p^f for $t - \tau$ (i.e., with peak at the inferred target), for $v = t - \tau + 1$ up to t. The resulting p^f is then normalized into a probability distribution. The z_{t+1} is then drawn randomly according to this distribution p^f. The only change compared to the generative algorithm presented before is in this update of the internal variable.

The likelihood model is changed along similar lines. In that case, the choice and reward are given, hence we need to consider $S(i, c_{t+1}, s_{t+1})p_i^f$, which measures the likelihood that the choice is made according to the updated p^f, which integrates the past τ trials. The new inferred target z_{t+1} is given by the feature i that maximizes this likelihood. The contribution to the objective function for trial $t + 1$ is given by the log likelihood $LL = log[\Sigma_i \, S(c_{t+1}, i, s_{t+1})p_i^f]$.

Examples

In Appendix Figure A2, we show the predictions of the generative model with random switching between features. A higher attentional load results in slower learning and a lower asymptotic performance (typically the experiment stops at 30, so asymptotic performance means performance toward the end of the block. In Appendix Figure A3, we show the predictions for feature-based switching. The key feature is that integrating over one previous trial is not enough to reach perfect performance for even the lowest attentional load, whereas for Load 3, none of the delays τ up to 9 is sufficient.

Acknowledgments

The authors thank Shelby Volden and Seth König for help with animal training and data collection and Seth König for preparing the data for analysis and for introducing the concept of a "neutral" Quaddle stimulus. Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number R01MH123687. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Reprint requests should be sent to Thilo Womelsdorf, Department of Psychology, Vanderbilt University, 301 Wilson Hall 111, 21st Avenue South, Nashville, TN 37240, or via e-mail: thilo.womelsdorf@vanderbilt.edu or Paul Tiesinga, Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen 6525 EN, The Netherlands, or via e-mail: P.Tiesinga@science.ru.nl.

Data and Code Accessibility

Data and computational modeling code for reproducing the results of the best-fitting model (Figure 4) is available at https://github.com/att-circ-contrl/Model-WM-RL-cooperation or from the corresponding authors.

Author Contributions

Thilo Womelsdorf: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing—Original draft; Writing—Review & editing. Marcus R. Watson: Data curation; Software; Writing—Review & editing. Paul Tiesinga: Conceptualization; Formal analysis; Methodology; Resources; Software; Supervision; Validation; Visualization; Writing—Original draft; Writing—Review & editing.

Funding Information

Thilo Womelsdorf, National Institute of Mental Health (https://dx.doi.org/10.13039/100000025), grant number: MH123687.

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., *JoCN*, 33:1, pp. 3–7). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

REFERENCES

Adams, R. P., & MacKay, D. J. C. (2007). Bayesian online changepoint detection. *arXiv:0710.3742*.

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

Alexander, W. H., & Brown, J. W. (2015). Hierarchical error representation: A computational model of anterior cingulate and dorsolateral prefrontal cortex. *Neural Computation*, *27*, 2354–2410. https://doi.org/10.1162/NECO_a_00779, PubMed: 26378874

Alexander, W. H., & Womelsdorf, T. (2021). Interactions of medial and lateral prefrontal cortex in hierarchical predictive coding. *Frontiers in Computational Neuroscience*, *15*, 605271. https://doi.org/10.3389/fncom.2021.605271, PubMed: 33613221

Averbeck, B. B. (2017). Amygdala and ventral striatum population codes implement multiple learning rates for reinforcement learning. In *2017 IEEE Symposium Series on*

Computational Intelligence (SSCI). Honolulu, HI: IEEE. https://doi.org/10.1109/SSCI.2017.8285354

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron*, *73*, 595–607. https://doi.org/10.1016/j.neuron.2011.12.025, PubMed: 22325209

Balcarras, M., Ardid, S., Kaping, D., Everling, S., & Womelsdorf, T. (2016). Attentional selection can be predicted by reinforcement learning of task-relevant stimulus features weighted by value-independent stickiness. *Journal of Cognitive Neuroscience*, *28*, 333–349. https://doi.org/10.1162/jocn_a_00894, PubMed: 26488586

Balleine, B. W. (2019). The meaning of behavior: Discriminating reflex and volition in the brain. *Neuron*, *104*, 47–62. https://doi.org/10.1016/j.neuron.2019.09.024, PubMed: 31600515

Banaie Boroujeni, K., Watson, M., & Womelsdorf, T. (2021). Gains and losses differentially regulate attentional efficacy at low and high attentional load. *bioRxiv*, 1–43. https://doi.org/10.1101/2020.09.01.278168

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*, 404–410. https://doi.org/10.1038/nn1209, PubMed: 15004564

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B: Methodological*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, *62*, 733–743. https://doi.org/10.1016/j.neuron.2009.05.014, PubMed: 19524531

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, *23*, 408–422. https://doi.org/10.1016/j.tics.2019.02.006, PubMed: 31003893

Cazé, R. D., & van den Meer, M. A. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics*, *107*, 711–719. https://doi.org/10.1007/s00422-013-0571-5, PubMed: 24085507

Chelazzi, L., Marini, F., Pascucci, D., & Turatto, M. (2019). Getting rid of visual distractors: The why, when, how, and where. *Current Opinion in Psychology*, *29*, 135–147. https://doi.org/10.1016/j.copsyc.2019.02.004, PubMed: 30856512

Collins, A. G. E., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *Journal of Neuroscience*, *34*, 13747–13756. https://doi.org/10.1523/JNEUROSCI.0989-14.2014, PubMed: 25297101

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working memory load strengthens reward prediction errors. *Journal of Neuroscience*, *37*, 4332–4342. https://doi.org/10.1523/JNEUROSCI.2700-16.2017, PubMed: 28320846

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*, 1024–1035. https://doi.org/doi.org/10.1111/j.1460-9568.2011.07980.x, PubMed: 22487033

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*, 190–229. https://doi.org/10.1037/a0030852, PubMed: 23356780

Dajani, D. R., & Uddin, L. Q. (2015). Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. *Trends in Neurosciences*, *38*, 571–578. https://doi.org/10.1016/j.tins.2015.07.003, PubMed: 26343956

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences, U.S.A.*, *95*, 14529–14534. https://doi.org/10.1073/pnas.95.24.14529, PubMed: 9826734

Donegan, N. H. (1981). Priming-produced facilitation or diminution of responding to a Pavlovian unconditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, *7*, 295–312. https://doi.org/10.1037/0097-7403.7.4.295, PubMed: 7288366

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*, 495–506. https://doi.org/10.1016/S0893-6080(02)00044-8, PubMed: 12371507

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: A model of attention in associative learning. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, *278*, 2553–2561. https://doi.org/10.1098/rspb.2011.0836, PubMed: 21653585

Failing, M., & Theeuwes, J. (2018). Selection history: How reward modulates selectivity of visual attention. *Psychonomic Bulletin & Review*, *25*, 514–538. https://doi.org/10.3758/s13423-017-1380-y, PubMed: 28986770

Farashahi, S., Donahue, C. H., Khorsand, P., Seo, H., Lee, D., & Soltani, A. (2017). Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. *Neuron*, *94*, 401–414. https://doi.org/10.1016/j.neuron.2017.03.044, PubMed: 28426971

Farashahi, S., Rowe, K., Aslami, Z., Lee, D., & Soltani, A. (2017). Feature-based learning improves adaptability without compromising precision. *Nature Communications*, *8*, 1768. https://doi.org/10.1038/s41467-017-01874-w, PubMed: 29170381

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences, U.S.A.*, *104*, 16311–16316. https://doi.org/10.1073/pnas.0706111104, PubMed: 17913879

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, *306*, 1940–1943. https://doi.org/10.1126/science.1102941, PubMed: 15528409

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101–128. https://doi.org/10.1146/annurev-psych-122414-033625, PubMed: 27618944

Hall, G., & Pearce, J. M. (1979). Latent inhibition of a CS during CS–US pairings. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*, 31–42. https://doi.org/10.1037/0097-7403.5.1.31, PubMed: 528877

Hassani, S. A., Oemisch, M., Balcarras, M., Westendorff, S., Ardid, S., van der Meer, M. A., et al. (2017). A computational psychiatry approach identifies how alpha-2A noradrenergic agonist guanfacine affects feature-based reinforcement learning in the macaque. *Scientific Reports*, *7*, 40606. https://doi.org/10.1038/srep40606, PubMed: 28091572

Ito, M., & Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience*, *29*, 9861–9874. https://doi.org/10.1523/JNEUROSCI.6157-08.2009, PubMed: 19657038

Kahnt, T., Park, S. Q., Cohen, M. X., Beck, A., Heinz, A., & Wrase, J. (2009). Dorsal striatal–midbrain connectivity in humans predicts how reinforcements are used to guide decisions. *Journal of Cognitive Neuroscience*, *21*, 1332–1345. https://doi.org/10.1162/jocn.2009.21092, PubMed: 18752410

Khamassi, M., Enel, P., Dominey, P. F., & Procyk, E. (2013). Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. *Progress in Brain Research*, *202*, 441–464. https://doi.org/10.1016/B978-0-444-62604-2.00022-8, PubMed: 23317844

Khamassi, M., Quilodran, R., Enel, P., Dominey, P. F., & Procyk, E. (2015). Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex. *Cerebral Cortex*, *25*, 3197–3218. https://doi.org/10.1093/cercor/bhu114, PubMed: 24904073

Klein, T. A., Neumann, J., Reuter, M., Hennig, J., von Cramon, D. Y., & Ullsperger, M. (2007). Genetically determined differences in learning from errors. *Science*, *318*, 1642–1645. https://doi.org/doi.org/10.1126/science.1145044, PubMed: 18063800

Kour, G., & Morris, G. (2019). Estimating attentional set-shifting dynamics in varying contextual bandits. *bioRxiv*, 621300. https://doi.org/10.1101/621300

Krugel, L. K., Biele, G., Mohr, P. N. C., Li, S.-C., & Heekeren, H. R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences, U.S.A.*, *106*, 17951–17956. https://doi.org/10.1073/pnas.0905191106, PubMed: 19822738

Kruschke, J. K. (2011). Models of attentional learning. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 120–152). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511921322.006

Lavie, N., & Fox, E. (2000). The role of perceptual load in negative priming. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 1038–1052. https://doi.org/10.1037/0096-1523.26.3.1038, PubMed: 10884008

Le Pelley, M. E., Pearson, D., Griffiths, O., & Beesley, T. (2015). When goals conflict with values: Counterproductive attentional and oculomotor capture by reward-related stimuli. *Journal of Experimental Psychology: General*, *144*, 158–171. https://doi.org/10.1037/xge0000037, PubMed: 25420117

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*, 0067. https://doi.org/10.1038/s41562-017-0067

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*, 451–463. https://doi.org/10.1016/j.neuron.2016.12.040, PubMed: 28103483

McDougle, S. D., & Collins, A. G. E. (2020). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, *28*, 20–39. https://doi.org/10.3758/s13423-020-01774-z, PubMed: 32710256

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, *126*, 292–311. https://doi.org/10.1037/rev0000120, PubMed: 30676040

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

Namburi, P., Beyeler, A., Yorozu, S., Calhoon, G. G., Halbert, S. A., Wichmann, R., et al. (2015). A circuit mechanism for differentiating positive and negative associations. *Nature*, *520*, 675–678. https://doi.org/10.1038/nature14366, PubMed: 25925480

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*, 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010, PubMed: 20844132

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., et al. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*, 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015, PubMed: 26019331

Noonan, M. P., Crittenden, B. M., Jensen, O., & Stokes, M. G. (2018). Selective inhibition of distracting input. *Behavioural Brain Research*, *355*, 36–47. https://doi.org/10.1016/j.bbr.2017.10.010, PubMed: 29042157

Oemisch, M., Westendorff, S., Azimi, M., Hassani, S. A., Ardid, S., Tiesinga, P., et al. (2019). Feature-specific prediction errors and surprise across macaque fronto-striatal circuits. *Nature Communications*, *10*, 176. https://doi.org/10.1038/s41467-018-08184-9, PubMed: 30635579

Papachristos, E. B., & Gallistel, C. R. (2006). Autoshaped head poking in the mouse: A quantitative analysis of the learning curve. *Journal of the Experimental Analysis of Behavior*, *85*, 293–308. https://doi.org/10.1901/jeab.2006.71-05, PubMed: 16776053

Pinherio, J. C., & Bates, D. M. (1996). Unconstrained parametrizations for variance–covariance matrices. *Statistics and Computing*, *6*, 289–296. https://doi.org/10.1007/BF00140873

Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia*, *41*, 245–251. https://doi.org/10.1016/S0028-3932(02)00157-4, PubMed: 12457750

Radulescu, A. (2020). *Computational mechanisms of selective attention during reinforcement learning*. Princeton, NJ: Princeton University.

Radulescu, A., Daniel, R., & Niv, Y. (2016). The effects of aging on the interaction between reinforcement learning and attention. *Psychology and Aging*, *31*, 747–757. https://doi.org/10.1037/pag0000112, PubMed: 27599017

Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic reinforcement learning: The role of structure and attention. *Trends in Cognitive Sciences*, *23*, 278–292. https://doi.org/10.1016/j.tics.2019.01.010, PubMed: 30824227

Rmus, M., McDougle, S. D., & Collins, A. G. E. (2020). The role of executive function in shaping reinforcement learning. *Current Opinion in Behavioral Sciences*, *38*, 66–73. https://doi.org/10.1016/j.cobeha.2020.10.003

Roelfsema, P. R., & Holtmaat, A. (2018). Control of synaptic plasticity in deep cortical networks. *Nature Reviews Neuroscience*, *19*, 166–180. https://doi.org/10.1038/nrn.2018.6, PubMed: 29449713

Rombouts, J. O., Bohte, S. M., & Roelfsema, P. R. (2015). How attention can create synaptic tags for the learning of working memories in sequential tasks. *PLoS Computational Biology*, *11*, e1004060. https://doi.org/10.1371/journal.pcbi.1004060, PubMed: 25742003

Rusz, D., Le Pelley, M. E., Kompier, M. A. J., Mait, L., & Bijleveld, E. (2020). Reward-driven distraction: A meta-analysis. *Psychological Bulletin*, *146*, 872–899. https://doi.org/10.1037/bul0000296, PubMed: 32686948

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, *27*, 12860–12867. https://doi.org/10.1523/JNEUROSCI.2496-07.2007, PubMed: 18032658

Schweighofer, N., & Arbib, M. A. (1998). A model of cerebellar metaplasticity. *Learning & Memory*, *4*, 421–428. https://doi.org/10.1101/lm.4.5.421, PubMed: 10701881

Seo, H., Cai, X., Donahue, C. H., & Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. *Science*, *346*, 340–343. https://doi.org/10.1126/science.1256254, PubMed: 25236468

Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, *20*, 635–644. https://doi.org/10.1038/s41583-019-0180-y, PubMed: 31147631

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge, MA: MIT Press.

Taswell, C. A., Costa, V. D., Murray, E. A., & Averbeck, B. B. (2018). Ventral striatum's role in learning from gains and losses. *Proceedings of the National Academy of Sciences, U.S.A.*, *115*, E12398–E12406. https://doi.org/10.1073/pnas.1809833115, PubMed: 30545910

Tomov, M. S., Truong, V. Q., Hundia, R. A., & Gershman, S. J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature Communications*, *11*, 2371. https://doi.org/10.1038/s41467-020-15766-z, PubMed: 32398675

van den Bos, W., Cohen, M. X., Kahnt, T., & Crone, E. A. (2012). Striatum–medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. *Cerebral Cortex*, *22*, 1247–1255. https://doi.org/10.1093/cercor/bhr198, PubMed: 21817091

van der Meer, M., Kurth-Nelson, Z., & Redish, A. D. (2012). Information processing in decision-making systems. *Neuroscientist*, *18*, 342–359. https://doi.org/10.1177/1073858411435128, PubMed: 22492194

Viejo, G., Girard, B., Procyk, E., & Khamassi, M. (2018). Adaptive coordination of working-memory and reinforcement learning in non-human primates performing a trial-and-error problem solving task. *Behavioural Brain Research*, *355*, 76–89. https://doi.org/10.1016/j.bbr.2017.09.030, PubMed: 29061387

Viejo, G., Khamassi, M., Brovelli, A., & Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in Behavioral Neuroscience*, *9*, 225. https://doi.org/10.3389/fnbeh.2015.00225, PubMed: 26379518

Voloh, B., Watson, M. R., König, S., & Womelsdorf, T. (2020). MAD saccade: Statistically robust saccade threshold estimation via the median absolute deviation. *Journal of Eye Movement Research*, *12*. https://doi.org/10.16910/jemr.12.8.3, PubMed: 33828776

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196. https://doi.org/10.3758/BF03206482, PubMed: 15117008

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., et al. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*, 860–868. https://doi.org/10.1038/s41593-018-0147-8, PubMed: 29760527

Watson, M. R., Voloh, B., Naghizadeh, M., & Womelsdorf, T. (2019). Quaddles: A multidimensional 3-D object set with parametrically controlled and customizable features. *Behavior Research Methods*, *51*, 2522–2532. https://doi.org/10.3758/s13428-018-1097-5, PubMed: 30088255

Watson, M. R., Voloh, B., Thomas, C., Hasan, A., & Womelsdorf, T. (2019). USE: An integrative suite for temporally-precise psychophysical experiments in virtual environments for human, nonhuman, and artificially intelligent agents. *Journal of Neuroscience Methods*, *326*, 108374. https://doi.org/10.1016/j.jneumeth.2019.108374, PubMed: 31351974

Westendorff, S., Kaping, D., Everling, S., & Womelsdorf, T. (2016). Prefrontal and anterior cingulate cortex neurons encode attentional targets even when they do not apparently bias behavior. *Journal of Neurophysiology*, *116*, 796–811. https://doi.org/10.1152/jn.00027.2016, PubMed: 27193317

Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. https://doi.org/10.7554/eLife.49547, PubMed: 31769410

Wilson, R. C., & Niv, Y. (2011). Inferring relevance in a changing world. *Frontiers in Human Neuroscience*, *5*, 189. https://doi.org/10.3389/fnhum.2011.00189, PubMed: 22291631

Womelsdorf, T., & Everling, S. (2015). Long-range attention networks: Circuit motifs underlying endogenously controlled stimulus selection. *Trends in Neurosciences*, *38*, 682–700. https://doi.org/10.1016/j.tins.2015.08.009, PubMed: 26549883

Womelsdorf, T., Thomas, C., Neumann, A., Watson, M. R., Boroujeni Banaie, K., Hassani, S. A., et al. (2021). A Kiosk Station for the assessment of multiple cognitive domains and cognitve enrichment of monkeys. *Frontiers in Behavioral Neuroscience*, *15*, 721069. https://doi.org/10.3389/fnbeh.2021.721069, PubMed: 34512289

Worthy, D. A., Otto, A. R., & Maddox, W. T. (2012). Working-memory load and temporal myopia in dynamic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1640–1658. https://doi.org/10.1037/a0028146, PubMed: 22545616