# Learning at variable attentional load requires cooperation between working memory, meta-learning and attention-augmented reinforcement learning

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

[1]*Department of Psychology, Vanderbilt University, Nashville, TN 37240*

[2] *School of Kinesiology and Health Science, Centre for Vision Research, York University, 4700 Keele Street, Toronto, Ontario M6J 1P3, Canada.*

[3]*Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen 6525 EN, Netherlands. Correspondence*

Abbreviated Title: **Flexible learning at varying attentional load**

**Corresponding Authors**: Thilo Womelsdorf (thilo.womelsdorf@vanderbilt.edu) and Paul Tiesinga (P.Tiesinga@science.ru.nl)

**Keywords:**

Cognitive flexibility, exploration-exploitation tradeoff, feature-based attention, nonhuman primate, reward prediction error

## Abstract

**Flexible learning of changing reward contingencies can be realized with different strategies. A fast learning strategy involves forming a working memory of rewarding experiences with objects to improve future 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 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 subjects is consistently best predicted with a model combining (*i*) fast working memory (*ii*) slower reinforcement learning from positive and negative prediction errors, as well as (*iii*) selective suppression of non-chosen features values, and (*iv*) meta-learning based adjustment of exploration rates given a memory trace of recent errors. These mechanisms cooperate differently at low and high attentional loads. While working memory was essential for efficient learning at lower attentional loads, learning from negative outcomes and meta-learning were essential for efficient learning at higher attentional loads. Together, these findings highlight that a canonical set of distinct learning mechanisms cooperate to optimize flexible learning when adjusting to environments with real-world attentional demands.**

## Introduction

Cognitive flexibility is realized through multiple mechanisms (Dajani and Uddin, 2015), including recognizing when the environmental demands change, the rapid updating of expectations and the shifting of response strategies to away from irrelevant towards 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 values depends on a fast learning mechanism. Previous studies suggest that such a fast learning strategy can be based on at least three key 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 working memory (Collins and Frank, 2012; Collins et al., 2014; Viejo et al., 2018). 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 (Gershman and Daw, 2017; Botvinick et al., 2019).

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

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 (Khamassi et al., 2015). Such a meta-learning strategy, for example, increases the rate of exploring options as opposed to exploiting previously learned value estimates (Tomov et al., 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 multi-arm bandit problems (Wang et al., 2018).

There is evidence for all three proposed strategies in learning, but only a few empirical studies characterize the contribution of different learning strategies. Thus, it is unknown whether working memory, attention-augmented reinforcement learning and meta-learning approaches are all used and cooperate during learning in differently complex environments.

To address this issue, we set out to systematically explore the contribution of various computational mechanisms for learning about the value 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. Using more distracting features increased attentional load, resulting in successively more difficult, slower learning behavior. We found that across monkeys, learning speed was best predicted by a computational RL model that combines working memory, attention augmented RL and meta-learning. The contribution of these individual learning mechanisms varied systematically with attentional load. WM contributed to learning speed particularly at low and medium load, meta-learning contributed at medium and high loads, while selective attention was an essential learning mechanism at all attentional loads.

## Results

**Behavioral performance**. We measured how six monkeys learned the relevance of object features in a learning task while varying the number of distracting, reward-irrelevant 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, in order to learn by trial-and-error which object feature predicted reward. The rewarded feature could be any one of 37 possible features values from 4 different feature dimensions (color, shape, pattern and arms) of multidimensional *Quaddle* objects (**Fig. 1A**) (Watson et al., 2019a). The rewarded feature stayed constant within blocks of 35-60 trials (**Fig. 1B**). Learning blocks varied in the number of non-rewarded, distracting features (**Fig. 1C**). Subjects had 5 sec to choose an object which triggered correct or error feedback (a yellow or cyan halo around the chosen object, **Fig. 1D**). The first of three experimental conditions was labeled 1-dimensional attentional load because all the distractor features were from the same dimension as the target feature (e.g. different body shapes, see examples in top row of **Fig. 1E**,**F**). At 2-dimensional 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-dimensional attentional load, object features varied along three dimensions (e.g. varying in body shapes, surface patterns, and arm styles) (bottom row in **Fig. 1E**,**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 1D, 2D, and 3D attentional load conditions, respectively. The number of trials in a block needed to learn the relevant feature, i.e. to reach 75% criterion performance increased for the 1D, 2D, and 3D attentional load condition from on average 6.5, to 13.5, and 20.8 trials (SE's: 4.2 / 8.3 / 6.9) (Kruskal-Wallis test, $p = 0.0152$, ranks: 4.8, 10.2, 13.6) (**Fig. 2A,B**). Learning did not differ for blocks with a rewarded feature of the same or of a different dimension as the rewarded feature in the immediately preceding block (intra- versus extradimensional block transitions; Wilcoxon Ranksum test, $p = 0.699$, ranksum = 36) (**Fig. 2C**).

Flexible learning can be influenced by target- and distractor- history effects (Le Pelley et al., 2015; Failing and Theeuwes, 2018; Chelazzi et al., 2019; Rusz et al., 2020), which may vary with attentional load. We first considered the possibility 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 than 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 = 0.156, signed rank = 3; **Fig. 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 = 0.0312, signed rank = 0; **Fig. 2D**, *right*), which was significantly more pronounced in the high (3D) than in the low (1D) attentional load condition (paired signed rank test, again: p = 0.0312, signed rank = 0; **Fig. 2E**). These learning history effects suggest that learned target features had a significant influence on future learning in our task, particularly at high attentional load, while learned distractors had only marginal or no effects on subsequent learning.

**Multi-component modeling of flexible learning of feature values.** To discern specific mechanisms underlying flexible feature-value learning in individual monkeys we fit a series of reinforcement learning (RL) models to their behavioral choices (*see* Material and Methods). These models formalized individual mechanisms and allowed characterizing their role in accounting for behavioral learning at varying attentional load. 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 prediction errors (RPEs) with a learning gain $\eta$:

$V_{i,t+1}^F = V_{i,t}^F + \eta(R_t - V_{i,t}^F)$ (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 $V_i$, $p_i^{RL} = \frac{\exp(\beta^{RL} \sum_{j \in O_i} V_j^F)}{\sum_j \exp(\beta^{RL} \sum_{k \in O_j} V_k^F)}$ (Sutton and Barto, 2018). These two mechanisms incorporate two learning parameters: the weighting of prediction error information by $\eta$ (often called the learning rate), $\eta(PE)$, and the degree to which subjects explore or exploit learned values represented by $\beta$, 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 (**Supplementary Table 1**). The first of these mechanisms enhanced the expected values of all object features that were chosen by decaying feature values of non-chosen objects. This selective decay improved the prediction of choices in reversal learning and probabilistic multidimensional feature learning tasks (Wilson and Niv, 2011; Niv et al., 2015; Radulescu et al., 2016; Hassani et al., 2017; Oemisch et al., 2019). It is implemented as decay $\omega^{RL}$ of feature values $V_i$ from non-chosen features and thereby enhanced the value estimate for chosen (and hence attended) features for the next trial *t*:

$$V_{i,t+1}^F = (1 - \omega^{RL}) V_{i,t}^F \qquad \text{(}\textit{Eq. 1}\text{: Decay of non-chosen feature values).}$$

As second mechanism we considered a working memory (WM) process that uploads the identity of rewarded objects in a short-term memory. Such a WM can improve learning of multiple stimulus-response mappings (Collins and Frank, 2012; Collins et al., 2017) and multiple reward locations (Viejo et al., 2015; Viejo et al., 2018). 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 $pChoice^{WM}$, which competes with the $pChoice^{RL}$ from the reinforcement learning component of the model. The actual behavioral choice is the weighted sum of the choice probabilities of the WM and the RL component $w(\,pChoice^{WM}) + (1-w)pChoice^{RL}$. A weight $w$ of >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 $\beta$ parameter of the standard RL formulation) when errors accumulate. Similar to (Khamassi et al., 2013) the mechanism uses an error trace $\beta_t^*$, which increases when a choice was not rewarded, by an amount proportional to the negative RPE for that choice with a negative gain parameter $\alpha_-$, and decreases after correct trials proportional to the positive RPE weighted by a positive gain parameter $\alpha_+$ (Khamassi et al., 2013):

$\beta_{t+1}^* = \beta_t^* + \alpha_+ \lfloor \delta_t \rfloor_+ - \alpha_- |-\delta_t|$ (*Eq. 2*: Adjustment of exploration rate),

where the RPE 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 $\beta_t^{RL,a} = \frac{\beta_{max}}{1+\exp\,(-\omega_1(\beta_t^* - \omega_2))}$ (Khamassi et al., 2013). Transiently increasing the exploration rate increases the chances to find relevant object features when there are no reliable, learned values as guide 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 $\alpha_-$ 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 (Badre et al., 2012; Balcarras et al., 2016). 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 (Leong et al., 2017; Oemisch et al., 2019). Finally, as a seventh parameter we separately modelled the weighting of negative PE's after error outcomes, $\eta_{Loss}$, and the weighting of positive PE's for correct outcomes, $\eta_{Gain}$, 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) (Frank et al., 2004; Frank et al., 2007; Kahnt et al., 2009; van den Bos et al., 2012; Caze and van der Meer, 2013; Lefebvre et al., 2017; Taswell et al., 2018). 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 cross-validated separately for each attentional load condition and for each individual monkey. We calculated the Akaike Information Criteria (AIC) to rank order the models according to how well they predicted actual learning behavior given the number of free parameters.

**Working memory, adaptive exploration and decaying distractor values supplement reinforcement learning.** We found that across monkeys and attentional load conditions the RL model that best predicted monkey's choices during learning had four non-standard components:

(*i*) working memory, (*ii*) non-chosen value decay, (*iii*) adaptive exploration rate, and (*iv*) a separate gain for negative RPE's ($\eta_{Loss}$) (**Fig. 3**). This model had the lowest Akaike Information Criterion on average across all six monkeys and was ranked first for four individual monkeys (Monkeys 1,3,4, and 5, **Fig 3A,B;** see **Supplementary Table 1** shows the complete list of free model parameters for the rank ordered models). It ranked third and fourth for the other two monkeys (**Fig 3A**). Other models ranked in the top 5 entailed a stickiness mechanism to account for perseverative choice tendencies (models ranked 2$^{nd}$ and 3$^{rd}$), used an adaptive exploration mechanism with a fixed parameter (that thereby reduced the number of free parameters, model ranked 4th), or lacked the adaptive exploration mechanism (models ranked 3$^{rd}$ and 5$^{th}$). The top-ranked, most-predictive model reproduced well the learning curves obtained from the monkeys (**Fig. 3C**).

To discern how the individual model mechanisms of the most predictive model contributed to the learning at low, medium and high attentional load, we simulated the choice probabilities for this full model and for partial models implementing only individual mechanisms of that full model separately for each load conditions (**Fig. 4A**,**B**). This analysis confirmed that the 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 ~5% (**Fig. 4C**). The reduced (partial) model that performed similarly well across all attentional loads used the decay of non-chosen features ($\omega^{RL}$) (ranked 16$^{th}$ among all models, **Fig. 4C**). All other partial models were performing differently at low and high attentional loads. The partial working memory model (with $\omega^{WM}$) predicted choices well for the 1D and 2D load conditions but failed to account for choices in the 3D load condition (**Fig. 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 2D and 3D load condition (**Fig. 4C**). The partial model with the separate weighting of negative prediction errors ($\eta_{Loss}$, ranked 25th, see **Fig. 3A**) showed overall better choice probabilities than the standard RL model (ranked 26th), but still failed in predicting 12-14% of the monkeys' choices when used as the only non-standard RL mechanisms (**Fig. 4C**).

To understand why WM was only beneficial at low and medium attentional load, but detrimental at high attentional load, 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 (**Fig. 5A**). Following a block switch the WM module uploaded an object as soon as it was rewarded and maintained that rewarded object in memory over only few trials. When the rewarded object was encountered again prior to decaying to zero it guided the choice of that object beyond what the RL module would have suggested (evident in trial six in **Fig. 5A-C**). This WM contribution is beneficial when the same object instance re-occurred within few trials, which happened more frequently with low and medium attentional load, but only rarely during high load. At this high load condition, the RL components are faithfully tracking the choice probability of the monkey, while the WM representation of recently rewarded objects is non-informative because (*1*) it can only make a small contribution as the number of stimuli in the block is much larger than the capacity and because (*2*) it does not remember rewarded objects long enough to be around when the objects are presented another time.

While 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

load (**Fig. 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 following the block switch, even when there are no competitive other features in terms of value, because the mechanism enhances exploring objects with previously non-chosen features. Our results suggest that such an adjustment of exploration can reduce the decline in performance at high attentional load (**Fig. 4C**), i.e. when subjects have to balance exploring the increased number of features with not losing track of already gained target information (**Fig. 5D**,**E**).

**The relative contribution of model mechanisms for learning and asymptotic performance.** The relative contributions of individual model mechanisms for different attentional loads can be inferred from their load-specific parameter values that best predicted monkey's learning when cross-validated for learning at each load separately (**Fig. 6**). WM was maintained longer for 2D than 1D (lower WM decay values), but at 3D load showed fast decay (higher $\omega^{WM}$ values) signifying that WM representations stopped contributing to learning at high load (**Fig. 6C**). When load increased the models showed a consistent decrease of the weighting of positive PE's ($\eta_{Gain}$ from ~0.15 to 0.1) and of the weighting of negative PE's ($\eta_{Loss}$, from ~0.6 to 0.4) (**Fig. 6E**, and **6G**). A potential explanation for the decrease in $\eta_{Gain}$ is that with more distracting features more trials are needed to determine what feature is the target, which is achieved with slower updating. The decay of non-chosen feature values ($\omega^{RL}$) was weaker with increased load across monkeys indicating a longer retention of values of non-chosen objects (**Fig. 6F**), which reflects protecting the target value when it is not part of the chosen (hence unrewarded) object. An event that occurs

more for high loads. Adaptive exploration rates ($\beta$*) increased on average from low to medium and high load (more negative values) signifying increased exploration after errors at higher load.

The parameter variations at increasing load could relate either to the learning speed or the asymptotic performance differences at different loads. To quantify their relative contributions, we correlated the values of each model parameter with a model-independent estimate of learning speed (number of trials to reach criterion performance), and with asymptotic accuracy (proportion of correct trials after learning criterion was reached). We found that values of three parameters significantly correlated with learning speed (**Fig. 7A**). Learning occurred significantly earlier (*i*) with larger prediction error weighting for rewarded trials ($\eta_{Gain}$, r = -0.69, p = 0.0008, FDR controlled at alpha 0.05), with higher prediction error weight for unrewarded trials ($\eta_{Loss}$, r = -0.62, p = 0.0031, FDR controlled at alpha 0.05), and (*iii*) with less decay (and thus better retention) of values of unchosen features ($\omega^{RL}$, r = -0.511, p = 0.0144, 1, FDR controlled at alpha 0.05) (**Fig. 7C**). The same three parameters also correlated significantly with the asymptotic performance that monkey's showed after learning was achieved ($\eta_{Gain}$: r = 0.7, p = 0.0006; $\eta_{Loss}$r = 0.48, p = 0.022; $\omega^{RL}$ , r = 0.53, p = 0.0118, 1; all FDR controlled at alpha 0.05). (**Fig. 7B**,**D**). Higher asymptotic performance was additionally correlated with lower exploration rates (realized by higher $\beta^{RL}$, r = 0.54, p = 0.0097, FDR controlled at alpha 0.05). As for the working memory components, we found that better asymptotic performance was linked to *less* influence of working memory of recently rewarded objects on current choices as reflected in a higher $\beta^{WM}$ value with better performance (r = 0.54, p = 0.0097, FDR controlled at alpha 0.05). In summary, better performance followed from a combination of choices that exploited value estimates less (lower $\beta^{RL}$), kept a

stronger value trace for non-chosen features (higher $\omega^{RL}$), and stronger weighting of prediction errors after correct and error outcomes ($\eta_{Gain}$, $\eta_{Loss}$).

**Model parameter values distinguish fast from slow learners.** We next tested which model parameters distinguished good from bad learners across attentional load conditions by sorting subjects according to their learning speed (trials-to-criterion). We then correlated the ranks for subjects 1 to 6 for each load condition with the learning speed (**Fig. 8**). Across all parameters the two learning weights ($\eta_{Gain}$, $\eta_{Loss}$) were significantly larger for subjects that learned faster ($\eta_{Gain}$: r = -049, p = 0.019; $\eta_{Loss}$: r = -0.51, p = 0.015; significance after control for FDR alpha = 0.05) (**Fig. 8C**,**D**). This finding illustrates that while multiple mechanism correlate with learning success (**Fig. 7**), the key distinction between good and bad learners lies in their updating speed after errors and after correct outcomes (**Fig. 8**).

## Discussion

We found that learning feature values under increasing attentional load is accounted for by a reinforcement learning framework that incorporates four non-standard RL mechanisms: (*i*) a value-decrementing mechanism that selectively reduces the feature values associated with the non-chosen object, (*ii*) a separate working memory module that retains representations of rewarded objects over a few trials, (*iii*) separate gains for enhancing values after positive prediction errors and for suppressing values after negative prediction errors, 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 were similarly important

for all six animals (**Fig. 3**), suggesting they constitute a canonical set of mechanisms underlying cognitive flexibility. Although subjects varied in how these mechanisms were weighted (**Fig. 6**) those with fast learning and hence high cognitive flexibility were distinguished by stronger weighting of positive and negative prediction errors. Taken together these results document a formally defined set of mechanisms that underlies flexible learning of feature relevance under variable attentional load.

**Selective value enhancement is a key mechanism to cope with high attentional load**. One key finding was that only one non-standard RL mechanism, the decay of values of non-chosen features ($\omega^{RL}$), contributed similarly to learning across all attentional load conditions, correlating with learning speed and with overall accuracy after learning (r=-0.62, r=0.48, respectively, **Fig 7C,D**). This finding highlights the importance of this mechanism and supports previous studies that used a similar decay of non-chosen features to account for learning in multidimensional environments with deterministic or probabilistic reward schedules (Wilson and Niv, 2011; Niv et al., 2015; Radulescu et al., 2016; Hassani et al., 2017; Oemisch et al., 2019). 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 a feature is chosen (or attended), its value is updated and contributes to the next choice, while 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 non-chosen stimulus features for learning (Hall and Pearce, 1979; Donegan, 1981; Esber and Haselgrove, 2011) or the

slowing of responses to those stimuli (also called negative priming) (Lavie and Fox, 2000). The effect is consistent with a plasticity process that selectively tags synapses of those neuronal connections that represent chosen objects in order to enable their plasticity while preventing (or disabling) plasticity of non-tagged synapses processing non-chosen objects (Rombouts et al., 2015; Roelfsema and Holtmaat, 2018). 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. Accordingly, only chosen objects are updated, resembling how $\omega^{RL}$ implements increasing values for chosen objects when rewarded and the passive decay of values of nonchosen objects. Consistent with this interpretation $\omega^{RL}$ was significantly positively correlated with learning speed (**Fig. 7A**). At low attentional load, high $\omega^{RL}$ values reflected fast forgetting of non-chosen stimuli, while at high attentional load $\omega^{RL}$ adjusted to lower values which slowed down the forgetting of values of nonchosen values. The lowering of the $\omega^{RL}$ decay at high load reflects that values of all stimulus features are retained in a form of choice-history trace. Consistent with this finding, various studies have reported that prefrontal cortex areas contain neurons representing values of unchosen objects and unattended features of objects (Boorman et al., 2009; Westendorff et al., 2016). Our results demonstrate that at high attentional load, 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 (**Fig. 7A**).

**Working memory supports learning in parallel to reinforcement learning**. We found empirical evidence that learning the relevance of visual features leverages a fast working memory mechanism in parallel with a slower reinforcement learning 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 (Poldrack and Packard, 2003; van der Meer et al., 2012; Balleine, 2019). 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 or switch response strategies (Worthy et al., 2012). A separate learning system is associative and implicitly integrates experiences over longer time periods (Poldrack and Packard, 2003), which in our model corresponds to the reinforcement learning module.

The WM mechanisms we adopted for the feature learning task is similar to WM mechanisms that contributed in previous studies to the learning of strategies of a matching pennies games (Seo et al., 2014), the learning of hierarchical task structures (Collins and Frank, 2012, 2013; Alexander and Brown, 2015), or the flexibly learning varying reward locations (Viejo et al., 2015; Viejo et al., 2018). Our study adds to these prior studies by documenting that the benefit of WM is restricted to tasks with low and medium attentional load. The failure of working memory to contribute to learning at higher load might reflect an inherent limit in working memory capacity. Beyond an interpretation that WM capacity limits are reached at higher load, WM is functionally predominantly used to facilitate processing of actively processed items as opposed to inhibiting the processing of items stored in working memory (Noonan et al., 2018). In other words, a useful working memory is rarely filled with distracting, non-relevant information that a subject avoids. In our task, high distractor load would thus overwhelm the working memory store with information about non-rewarded objects whose active use would lead to no reward. Consequently, the model

– and the subject whose choices the model predicts – downregulate the importance of WM at high attentional load, relying instead on slower reinforcement learning 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 prediction errors improved model predictions of learning across attentional loads (**Fig. 4B**) by allowing a ~3-fold larger learning rate for negative than positive outcomes. 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 tasks (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 (Caze and van der Meer, 2013). According to this rationale, larger learning rates for negative outcomes in our task promotes 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 non-adaptive as it would promote frequent switching choices which is suboptimal in these probabilistic environments (Caze and van der Meer, 2013). These

considerations can also explain why multiple prior studies with probabilistic reward schedules report of an overweighting of positive over negative prediction errors which in their tasks promoted staying with and prevent switching from recent choices (Frank et al., 2004; Frank et al., 2007; Kahnt et al., 2009; van den Bos et al., 2012; Lefebvre et al., 2017).

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 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 closely associated with the ventral striatum (Namburi et al., 2015; Averbeck, 2017; Taswell et al., 2018). Our finding of considerably higher (faster) learning rates for negative than positive prediction errors is consistent with this view of a fast versus a 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 fro both, positive and negative outcomes, distinguished good and bad learners (**Fig. 8**), which supports reports that good and bad learning human subjects differ prominently in their strength of prediction error updating signals (Schonberg et al., 2007).

**Conclusion**

Our study provides quantitative evidence that a pure reinforcement learning modeling approach does not capture the cognitive processes needed to solve feature-based learning. By formalizing

the subcomponent processes needed to augment standard 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 et al., 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 working memory, adaptive exploration, and an ‘attention’ mechanism for decaying nonchosen values. As outlined in **Fig. 9** these three mechanisms leverage distinct learning signals, updating values based directly on outcomes (WM), on prediction errors (RL based decay of nonchosen values), or on a continuous error history trace (meta-learning based adaptive exploration). As a consequence, these three learning mechanisms operate in parallel with a relative influence on the decision that varies with task conditions, for instance, learning fast versus slow (WM versus Meta-learning versus RL) and adapting optimally to low versus high attentional load (WM versus Meta-learning). We believe that our empirical documentation of these mechanisms in monkeys learning the relevance of features is an important starting point to identify how neural systems integrate these mechanisms during cognitive flexible behavior.

## Materials and Methods

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

The experiment was controlled by USE (Unified Suite for Experiments) using the Unity 3D game engine for behavioral control and visual display (Watson et al., 2019b). Four animals performed the experiment in cage-based touchscreen Kiosk Testing Station described in (Womelsdorf et al., in preparation), while two animals performed the experiment in a sound attenuating experimental booth. All experiments used three dimensionally rendered objects, so called Quaddles (Watson et al., 2019a), that were defined by their body shape, arm style, surface pattern, and color (**Fig. 1A**). We used up to nine possible body shapes, six possible colors, eleven 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 ~3cm on the screen corresponding to ~2.5° degrees of visual angle and were presented either on a 24'' BenQ monitor or an Elo 2094L 19.5 LCD touchscreen running at 60 Hz refresh rate with 1920 x 1080 pixel resolution.

**Task paradigm.** Animals performed a feature-based choice task that required learning through trial-and-error which feature of multidimensional objects is associated with reward. The feature that was reward associated stayed constant for blocks of 35-60 trials and then switched randomly to another feature (**Fig. 1B**). Individual trials (**Fig. 1C**) were initiated by either touching a central

blue square (4 monkeys) or fixating the blue square for 0.5 sec. Following a 0.3 sec. delay three objects were presented at random locations of a 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). Following the choice of an object visual feedback was provided as a colored disk behind the selected object (yellow/grey for rewarded/not rewarded choices, respectively) concomitant with auditory feedback (low/high pitched sound for non-rewarded/rewarded choices, respectively). Choices of the object with the rewarded feature resulted in fluid reward at 0.3 sec. following 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 non-varying features were always either a spherical body shape, straight arms with blunt ending, grey color and 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 non-neutral feature values rendering them 1-, 2-, and 3- dimensional (**Fig. 1D**). For blocks with objects that varied in one feature dimension (1D attentional load condition) three feature values from that dimensions were chosen at the beginning of the block (e.g. body shapes that were oblong, pyramidal, and cubic). One of these features were associated with reward while 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 successive example trials in **Fig. 1E,F** (*upper* 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 while the other two feature values of that dimension and the feature values of the other dimension were not linked to reward. **Fig. 1E,F** (*middle* row) illustrates three example trials of these blocks of the 2D attentional load condition. For blocks with objects varying in three feature dimensions (3D 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 twenty-seven unique objects combining these features. Again, only one feature was associated with reward while all other feature values were not linked to reward.

Blocks with objects that varied in 1, 2 and 3 feature dimensions constitute 1D, 2D, and 3D 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 the dimensionality of the learning problem varied pseudo-randomly 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 600 Hz sampling rate and 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 elsewhere (Voloh et al., 2020).

**Data analysis**. All analysis was performed with custom MATLAB code (Mathworks, Natick, MA). Significance tests control for the false discover rate (FDR) with an alpha value of 0.05 to account for multiple comparisons (Benjamini and Hochberg, 1995).

**General formulation of Rescorla-Wagner Reinforcement Learning Models**. The value of feature *i* in trial *t*, before the outcome was known, is denoted by $V_{i,t}^{F}$. The superscript F stands for feature, in order to distinguish it from the value of an object that will be introduced in the next section. The new value, $V_{i,t+1}^{F}$ available for decisions on the next trial depends on which features were at t+1 present in the chosen object, and whether this choice was rewarded $R_t$ = 1, or not $R_t$ = 0. The value of features that were present in objects that were not chosen, and those that could appear in the course of the session, but were not present on the current trial, decay with a decay constant $\omega_t^{RL}$, where RL denotes that the decay component is from the reinforcement component of the model as opposed to decay of the working memory introduced below. The features that were present in the chosen and rewarded object increase in value, because the reward prediction error, $R_t - V_{i,t}^{F}$ is positive, whereas when the chosen object was not rewarded, the value decays. We have summarized these update rules in the following equations:

$$V_{i,t+1}^{F} = V_{i,t}^{F} + \eta_t \, f_{i,t}^{A,V}(R_t - V_{i,t}^{F}) \,, \qquad \text{features of objects} \qquad \text{(eq. 3)}$$

$$= (1 - \omega_{nc}^{RL}) V_{i,t}^{F} \qquad \text{features of non-chosen objects} \qquad \text{(eq. 4)}$$

$$= (1 - \omega_{np}^{RL}) V_{i,t}^{F} \qquad \text{non-presented features} \qquad \text{(eq. 5)}$$

We have indicated a trial-dependence in gain $\eta$ and allow the decay constants ω to depend on whether the feature was present in the non-chosen object (nc) or whether it as 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 reinforcement learning formulation rather than working memory (superscript WM). The setting of these parameters depends on the specific model version. In the base RL model, there is no feature-value decay $\omega_{nc,t} = \omega_{np,t} = 0$ 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

$\eta t = \eta_{Gain}$ Rt + $\eta_{Loss}$ (1 − Rt), which introduces two new parameters $\eta_{Gain}$ and $\eta_{Loss}$ for rewarded and non-rewarded choices, respectively.

In most models the decay for non-chosen and not presented features was equal $\omega nc,t = \omega np,t = \omega^{RL}$, introducing only a single additional parameter. In the so-called hybrid models we add a feature dimension gain factor $f_{i,t}^{A,V}$ , which reflects attention to a particular dimension. It is calculated using a Bayesian model, 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 $p_{i,t}^{RL}$ for object *i* at trial *t* is determined using a softmax function:

$$p_{i,t}^{RL} = \frac{\exp\left(\beta_t \sum_{j \in O_i} f_{j,t}^{A,CP} V_{j,t}^{F}\right)}{\sum_i \exp\left(\beta_t \sum_{j \in O_i} f_{j,t}^{A,CP} V_{j,t}^{F}\right)} \qquad \text{(eq. 6)}$$

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 $\beta_t$ in the exponent determines the extent to which the subject exploits, i.e. systematically chooses the object with the highest compound value, (reflected in a large β) or explores, i.e. makes choices independent of the compound value (reflected in small β). In most model versions the β did not change over trials, while 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.

**Adaptive Exploration.** For models with adaptive $\beta_t$ values we follow the model of (Khamassi et al., 2013), which involves determining an error trace:

$$\beta_{t+1}^{*} = \beta_t^{*} + \alpha_{+} max(\delta_t, 0) + +\alpha_{-} min(\delta_t, 0) \qquad \text{(eq. 7)}$$

where the min and max functions are used to select the negative and positive part, respectively, of an estimate of the reward prediction error,

$$\delta_t = R_t - \frac{1}{\#(j \in O_i)} \sum_{j \in O_i} V_{j,t}^F \qquad \text{(eq. 8)}$$

This is a different form of the RPE 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

$$\beta_t = \frac{\beta_m}{1 + exp\left(-\omega_1(\beta_t^* - \omega_2)\right)} \qquad \text{(eq. 9)}$$

This adaptive sub model replaces one parameter by five new parameters: $\alpha+$, $\alpha-$, $\beta m$, $\omega_1$ and $\omega_2$, of which we fixed four in most models, $\alpha+ = -0.6$, $\alpha- = -0.4$, $\omega_1 = -6$, $\omega_2 = 0.5$, and varied $\beta_m$ and sometimes varied $\alpha-$ as well.

**Attentional dimension weight.** The attentional gain factor 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 the target dimension (Niv et al., 2015; Hassani et al., 2017; Oemisch et al., 2019). From the target feature probability $p(f|\mathcal{D}_{1:t})$ we can obtain a target dimension probability by summing over all the feature values $f$(d) that belong to a particular dimension d,

$$p_{d,t}^D = p(d|\mathcal{D}_{1:t}) = \sum_{f \in f(d)} p(f|\mathcal{D}_{1:t}) \qquad \text{(eq. 10)}$$

this is turned into a feature gain

$$\phi_{d,t}^A = \frac{\left(p_{d,t}^D\right)^\alpha}{\sum_e \left(p_{e,t}^D\right)^\alpha} \qquad \text{(eq. 11)}$$

which weighs feature values in each object according to their dimension d(f), for an object *i*, this becomes $V_{i,t} = \sum_{j \in O} \phi_{d(j),t}^A V_{j,t}^F$, and which we incorporate as feature dependent factor as $f_{i,t}^A =$

$\phi^{A}_{d(j),t}$ in the relevant expressions (eq. (3)) and (6), indicated with additional superscript V and CP, respectively).

**Stickiness in Choice Probability.** Stickiness in object choice refers to choosing the object whose feature values overlap with the previously chosen one and represents a kind of perseverance (Balcarras et al., 2016). It is implemented by making the choice probability dependent on whether a feature on the previous trial is present in an object.

$$p_i^S = \frac{\exp\left(\beta_t \sum_{j \in O_i} f_{j,t}^{A,CP} V_{j,t}^{F}\right) + \Delta_{t-1}}{\sum_i \exp\left(\beta_t \sum_{j \in O_i} f_{j,t}^{A,CP} V_{j,t}^{F}\right) + \Delta_{t-1}} \qquad \text{(eq. 12)}$$

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 one, we assure 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 chosen object.

***Combined Working Memory – Reinforcement models.*** Working memory models are formulated in terms of the value $V_{i,t}^{WM}$ of an object irrespective of what features are present in it (Collins and Frank, 2012). These values are initialized to a non-informative 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 $V_{i,t+1}^{WM} = 1$ when rewarded, whereas it is reset to the original value $V_{i,t+1}^{WM} = \frac{1}{n_o}$ when the choice was not rewarded. All other values decay towards the original value with a decay constant $\omega^{WM}$:

$$V_{i,t+1}^{WM} = V_{i,t}^{WM} - \omega^{WM}\left(V_{i,t}^{WM} - \frac{1}{n_o}\right) \qquad \text{(eq. 13)}$$

The values are then directly used in the choice probabilities:

$$p_{i,t}^{WM} = \frac{\exp\,(\beta^{WM} V_{i,t}^{WM})}{\sum_j \exp\,(\beta^{WM} V_{j,t}^{WM})} \qquad \text{(eq. 14)}$$

This component mechanism thus introduces two new parameters, a decay constant $\omega^{WM}$ and $\beta^{WM}$, 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),

$$p_{i,t}^{T} = w_t p_{i,t}^{WM} + (1 - w_t) p_{i,t}^{RL} \qquad \text{(eq. 15)}$$

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 working memory:

$$w_{t+1} = \frac{p_t^{WMC} w_t}{w_t p_t^{WMC} + (1 - w_t) p_t^{RLC}} \qquad \text{(eq. 16)}$$

where

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

This expression selects from amongst 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,

$$p_t^{WMC} = \alpha\left(p_{a(t),t}^{WM} r_t + (1 - p_{a(t),t}^{WM})(1 - r_t)\right) + (1 - \alpha)\left(\frac{1}{n_o}\right) \qquad \text{(eq. 17),}$$

where $\alpha = \min\left(1, \frac{C_{WM}}{n_S}\right)$ and $C_{WM}$ is the working memory capacity, essentially the number of objects about which information can be accessed, $n_S$ is the number of objects that can be presented during the task. It is determined as the number of objects whose value $V_{i,t}^{WM}$ exceeds $\frac{1}{n_o}$ 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, can not be read out, instead $p_t^{WMC} = 1/n_o$ . Hence, when $p_{a(t),t}^{RL}$ exceeds $\frac{1}{n_o}$ it will win the competition for influence and reduce $w_t$ towards WM zero, and with that the influence of $p_{i,t}^{WM}$.

**Cross-validation**. We used a cross-validation procedure for evaluating how well the model predicted the subject's choices of (test-) learning blocks that were withheld when fitting the model parameters on the remaining (training-) datasets. 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 set. 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. These model simulations provided the log likelihood (normalized by the number of choices) that reflect how well the model reproduces the monkey's pattern of correct and erroneous choices. To compare models, we rank-ordered the Akaike Information Criterion (AIC) for each model that penalizes models according to the number of free parameters.

## Data and code availability

All data supporting this study and its findings, as well as custom MATLAB code generated for analyses, are available from the corresponding author upon reasonable request.


## Acknowledgements

The authors would like to 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. This work was supported by a grant from the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (R01EB028161). The funder had no role in study design, data collection and analysis, the decision to publish, or the preparation of this manuscript.


## Author Contributions

T.W. and M.R.W. conceived the experiments. M.R.W. wrote the code to run the experiments. T.W. and P.T. analyzed the data. T.W., M.R.W., and P. T. wrote the paper.

## Competing interests

The authors declare no competing interests.

## Figure legends

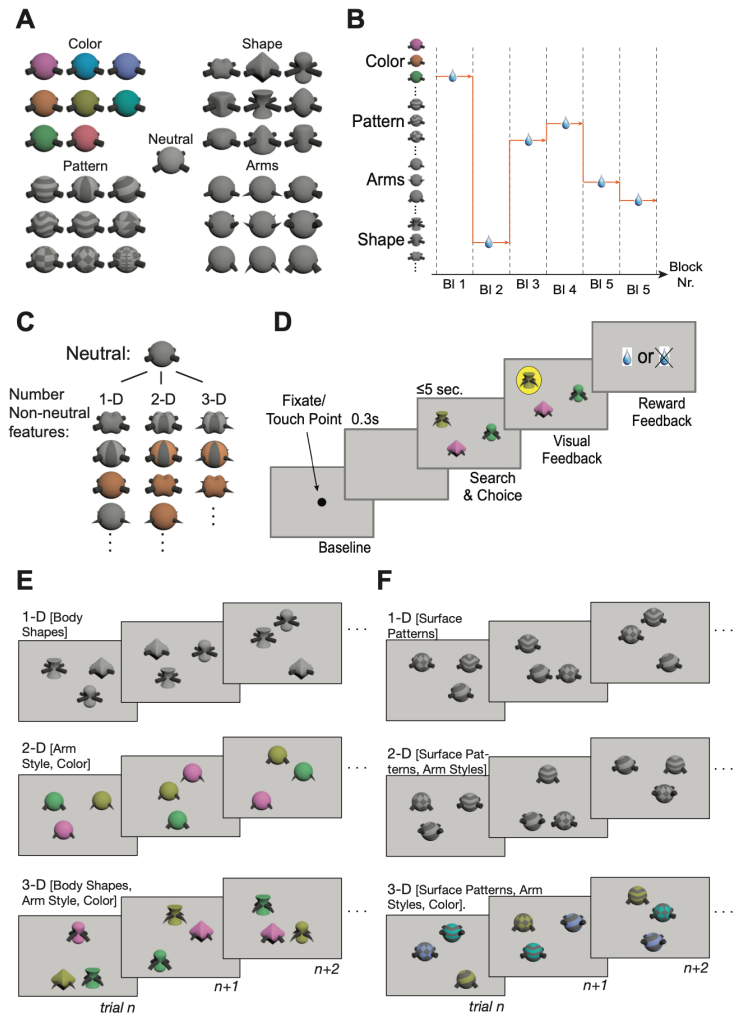


**Figure 1. Task paradigm and feature space.** (**A**) The task used 3D rendered 'Quaddle' objects that varied in color, pattern, shape and arm style. The features grey color, straight arms and spherical body shape were never rewarded in any of the experiments and therefore constitute 'neutral' features. (**B**) For successive blocks of 30-60 trials a single feature was rewarded. (**C**) The attentional load conditions differed in the number of non-neutral feature dimensions that varied across trials in a block. Blocks with 1-D, 2-D, and 3-D objects contained stimuli varying features in 1, 2, and 3 features 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 an object (two monkeys) for ≥0.7 sec. Visual feedback indicated correct (yellow) vs error (cyan, not shown) outcomes. Fluid reward followed correct outcomes. (**E**) Sequences of three example trials for a block with 1-D objects (upper row, shape varied), 2-D objects (middle, color and arms varied), and 3-D objects (bottom row, body shape, arms and color) objects. (**F**) Same as *E* but for an object set varying surface pattern, arms, and color.

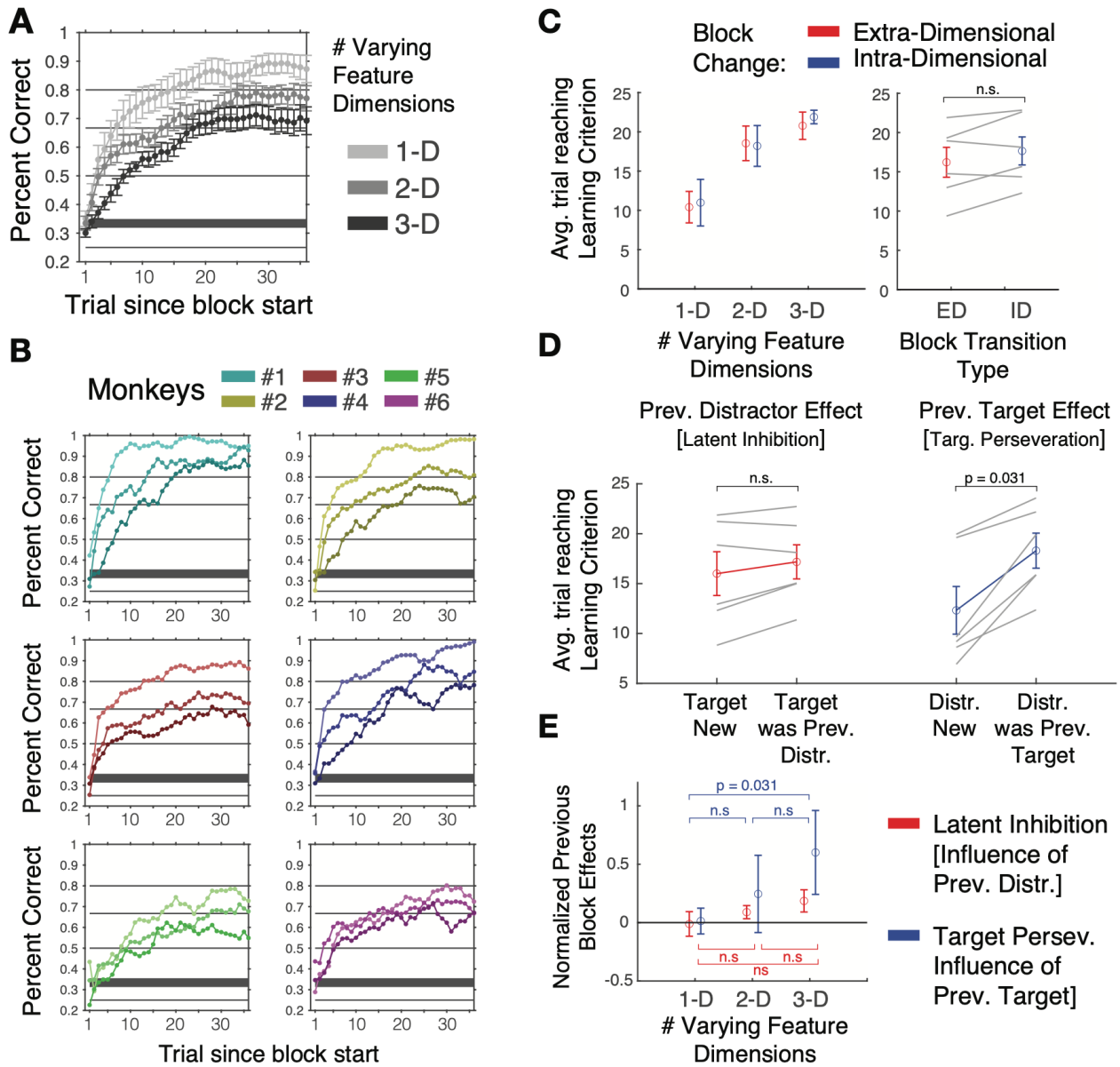


**Figure 2. Learning performance.** (**A**) Average Learning curves across six monkeys for the 1D, 2D, and 3D load condition. (**B**) Learning curves for each monkeys (colors) for 1D, 2D, 3D (low-to-high color saturation levels). All monkeys showed fastest learning for low load and slowest learning for the high load condition. Curves are smoothed with a 5 trial forward-looking window. (**C**) *Left:* The average trials-to-criterion (75% accuracy over 10 consecutive trials) for low to high attentional load (*x-axis*) for blocks in which the target feature was either of the same - intra-dimensional (ID) - or different dimension - extra-dimensional (ED) - as in the preceding trial. *Right*: Average number of trials-to-criterion across load conditions. Grey lines denote individual monkeys. Errors are SE. (**D**) The red color dneotes average trials-to-criterion for blocks in which the target feature was novel (not shown in 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 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 load. Errors are SE.

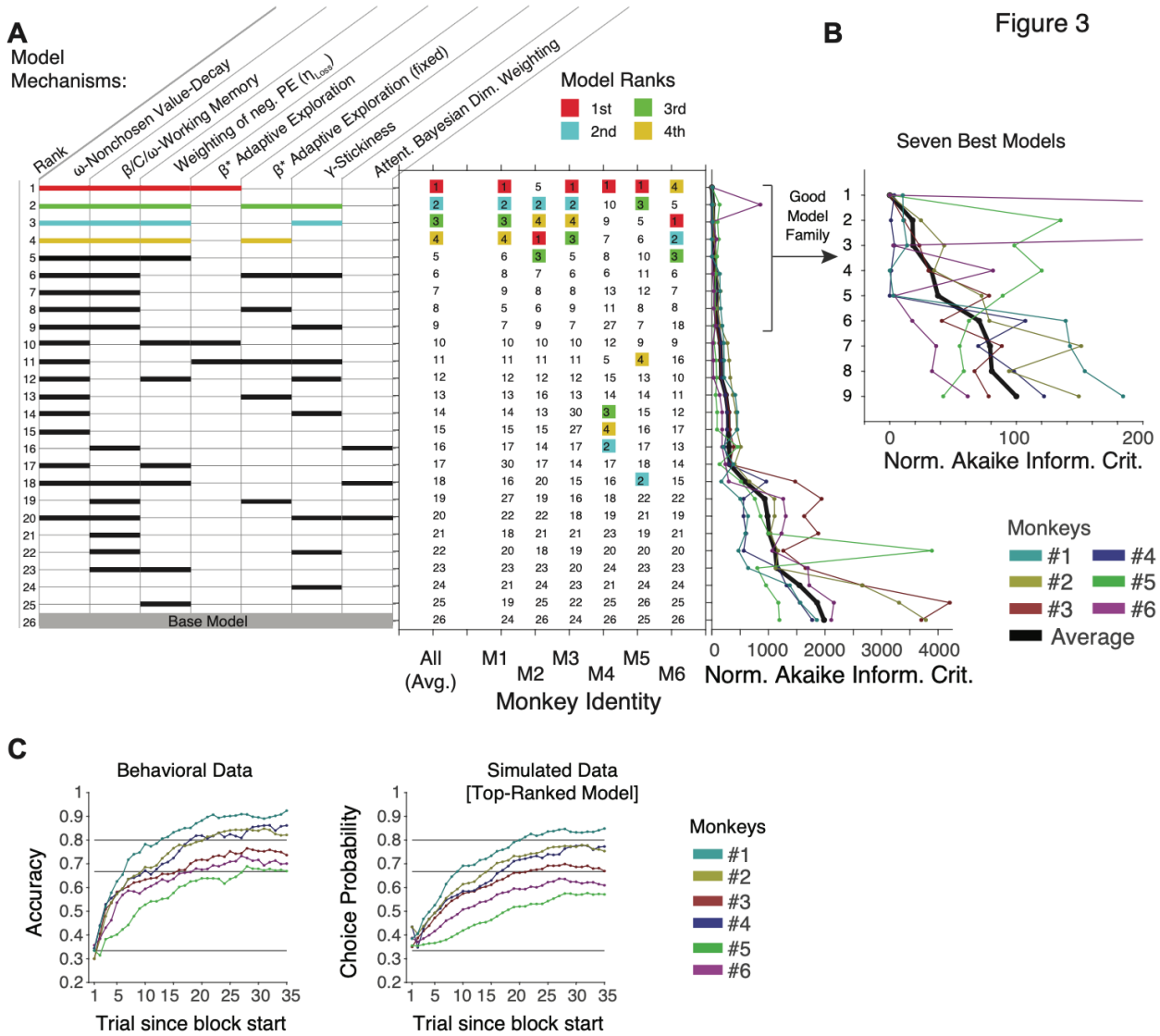


**Figure 3. Rank ordering of models with different combinations of mechanisms.** (**A**) Models (*rows*) using different combinations of model mechanisms (*columns*) are rank ordered according to their Akaike Information Criterion (AIC). The top-ranked model combined four mechanisms that are highlight in red: Decay of non-chosen features, working memory, 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. Thick horizonal 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. *Right*: Model rank average (1st column) and for each individual monkey (columns 2 to 7). See **Supplementary 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 AIC of the 1st ranked model, the normalized AIC's for each monkey confirms that the top-ranked model has low AIC 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.

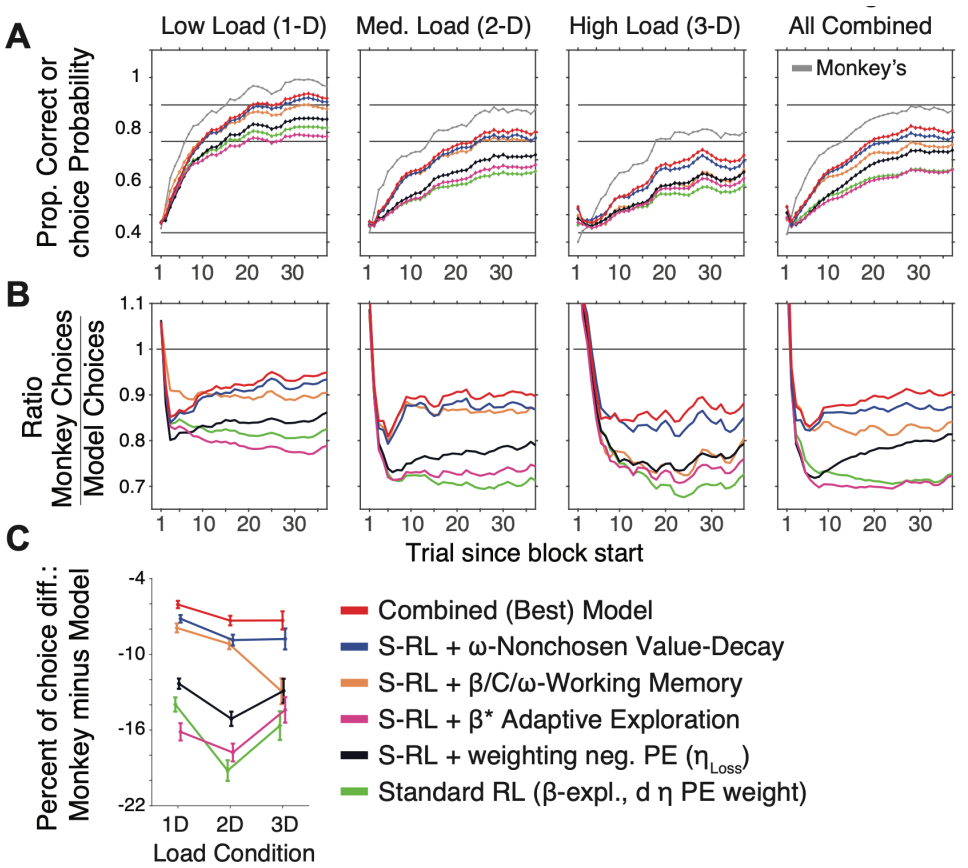


**Figure 4. Choice probabilities of monkeys and models at three different loads.** (**A**) Average choice accuracy of monkeys (grey) and choice probabilities of six models. The top-ranked model (*red*) combines WM with RL and selective suppression of nonchosen values, a separate learning gain for neg. RPE's, 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 are SE.

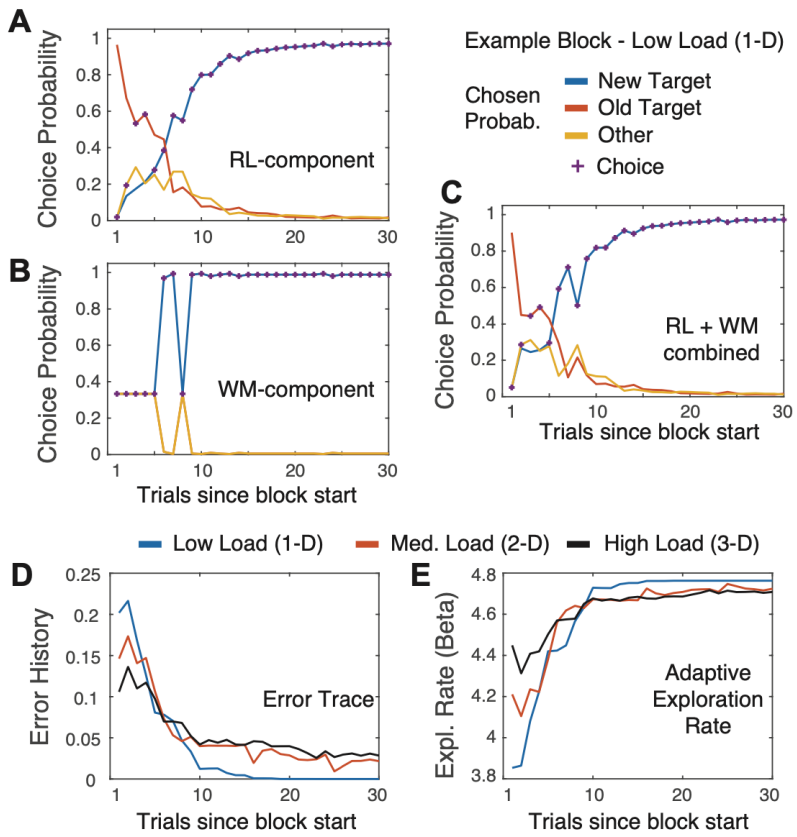


**Figure 5. Contribution of working memory, reinforcement learning and adaptive exploration to learning behavior.** (**A**) Choice probabilities of the RL component of the top-ranked model for an example block, calculated for the objects with the new target feature (*blue*), the previous block's target 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 working memory module of the model. Choice probabilities of the WM and the RL component 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 influences 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 non-zero 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.

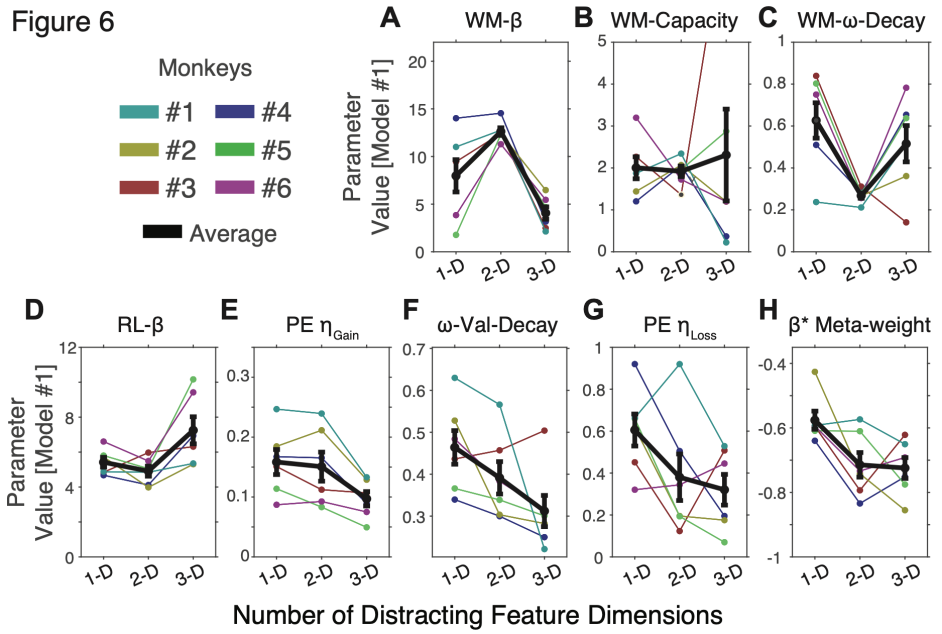


**Figure 6. Model parameter values at different attentional loads.** The average parameter values (*black*) of the top-ranked model (*y-axis*) plotted against the number of distracting feature dimensions for the WM parameters (**A-C**) and the RL parameters (**D-H**). Individual monkeys are in colors.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

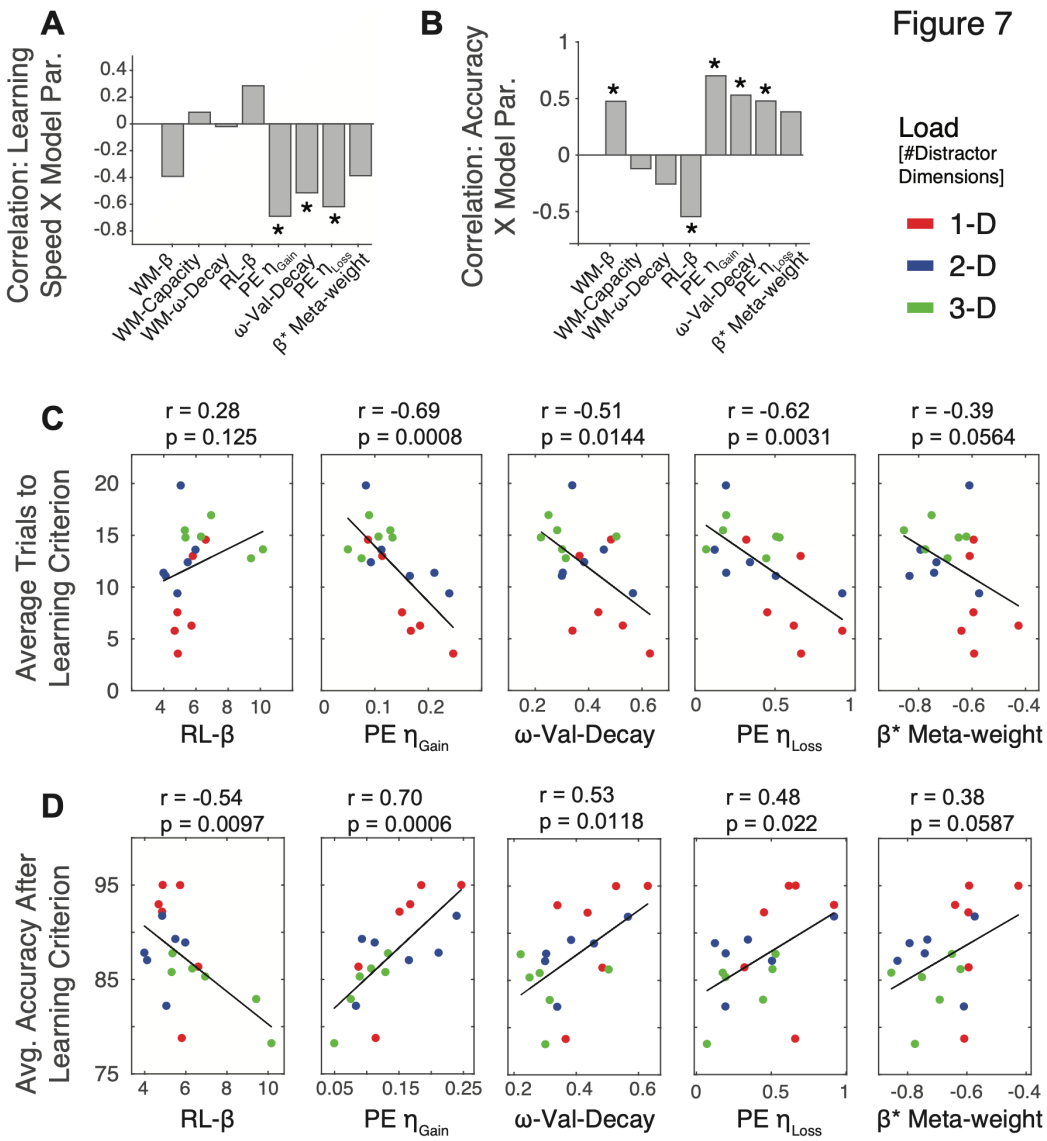


**Figure 7. Model parameter values underlying learning speed and asymptotic performance.** (**A**) The correlation across monkeys and attentional load conditions of learning (trials-to-criterion, i.e. less trials signify faster learning) and parameter values of the top-ranked model. Stars denote FDR corrected significance at $p < 0.05$. Negative correlations denote that higher parameter values associate with faster learning. (**B**) As A but for correlations of monkey's asymptotic performance accuracy with model parameter values. (**C,D**) Illustration of the correlations shown in *A,B* for the RL parameters. Y-axis shows the learning speed (trial to criterion) and x-axis the parameter values of the top-ranked model. The black line is the regression whose r and p values are given above each plot. Each dot is the average result from one monkey in either the 1D (red), 2-D (blue) or 3-D (green) condition.

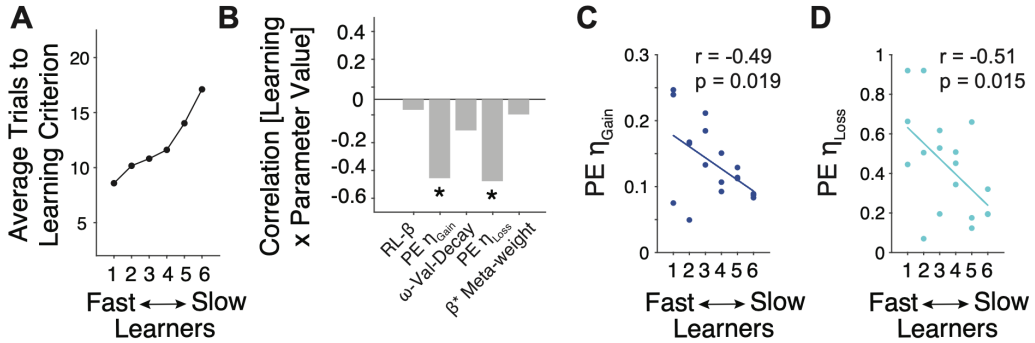


**Figure 8. Model mechanisms distinguishing slow and fast learners.** (**A**) The average learning speed (the trials to reach criterion, *y-axis*) plotted against the individual monkeys ordered from fastest to slowest learner. (**B**) Correlation of learning speed (trials-to-criterion) and rank-ordered monkeys (ranks #1 to #6). Stars denote FDR corrected significance at $p < 0.05$. (**C**,**D**) The weighting of positive prediction errors (*C*) and negative prediction errors (*D*) of the top-ranked model was significantly larger for fast than for slow learners.

| | Working Memory | 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 working memory, reinforcement learning and meta-learning components.** The model components differ in the teaching signals that trigger adjustment, in the learning speed and in how important they are to contribute to learning at increasing attentional load.

**Supplementary Table 1.** Overview of the parameter used in models evaluated and ranked according to AIC higher than the base RL model (which is the model ranked 26th). See figure 3 of the main text for a graphical illustration of the model rank ordering.

| Model Rank | RL gain ($\eta$) | RL decay ($\omega$) | RL softmax ($\beta$) | RL atn ($\alpha$) | WM decay ($\omega$) | WM softmax ($\beta$) | WM capacity ($C_{WM}$) | stickiness ($\gamma$) | #para |
|---|---|---|---|---|---|---|---|---|---|
| 1 | $\eta_{Gain}$,$\eta_{Loss}$ | $\omega^{RL}$ | $\beta_m$, $\alpha_-$ ($\alpha_+$, $\omega_1$, $\omega_2$ fix ) | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 8 |
| 2 | $\eta_{Gain}$,$\eta_{Loss}$ | $\omega^{RL}$ | $\beta_m$ ($\alpha_+$,$\alpha_-$,$\omega_1$,$\omega_2$fix) | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | $\gamma$ | 8 |
| 3 | $\eta_{Gain}$,$\eta_{Loss}$ | $\omega^{RL}$ | $\beta^{RL}$ | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | $\gamma$ | 8 |
| 4 | $\eta_{Gain}$,$\eta_{Loss}$ | $\omega^{RL}$ | $\beta_m$ ($\alpha_+$,$\alpha_-$,$\omega_1$,$\omega_2$fix) | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 7 |
| 5 | $\eta_{Gain}$,$\eta_{Loss}$ | $\omega^{RL}$ | $\beta^{RL}$ | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 7 |
| 6 | $\eta$ | $\omega^{RL}$ | $\beta_m$ ($\alpha_+$,$\alpha_-$,$\omega_1$,$\omega_2$fix) | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 7 |
| 7 | $\eta$ | $\omega^{RL}$ | $\beta^{RL}$ | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 6 |
| 8 | $\eta$ | $\omega^{RL}$ | $\beta_m$ ($\alpha_+$,$\alpha_-$,$\omega_1$,$\omega_2$fix) | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 6 |
| 9 | $\eta$ | $\omega^{RL}$ | $\beta_m$, $\alpha_-$ ($\alpha_+$, $\omega_1$, $\omega_2$ fix ) | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | $\gamma$ | 7 |
| 10 | $\eta_{Gain}$,$\eta_{Loss}$ | $\omega^{RL}$ | $\beta_m$, $\alpha_-$ ($\alpha_+$, $\omega_1$, $\omega_2$ fix ) | - | - | - | - | - | 5 |
| 11 | $\eta_{Gain}$,$\eta_{Loss}$ | $\omega^{RL}$ | $\beta_m$ ($\alpha_+$,$\alpha_-$,$\omega_1$,$\omega_2$fix) | - | - | - | - | $\gamma$ | 5 |
| 12 | $\eta_{Gain}$,$\eta_{Loss}$ | $\omega^{RL}$ | $\beta^{RL}$ | - | - | - | - | $\gamma$ | 5 |
| 13 | $\eta$ | $\omega^{RL}$ | $\beta_m$ ($\alpha_+$,$\alpha_-$,$\omega_1$,$\omega_2$fix) | - | - | - | - | - | 3 |
| 14 | $\eta$ | $\omega^{RL}$ | $\beta^{RL}$ | - | - | - | - | $\gamma$ | 4 |
| 15 | $\eta$ | $\omega^{RL}$ | $\beta^{RL}$ | - | - | - | - | - | 3 |
| 16 | $\eta$ | $\omega^{RL}$ | $\beta^{RL}$ | $\alpha$ | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 7 |
| 17 | $\eta_{Gain}$,$\eta_{Loss}$ | $\omega^{RL}$ | $\beta^{RL}$ | - | | | | - | 4 |
| 18 | $\eta_{Gain}$,$\eta_{Loss}$ | $\omega^{RL}$ | $\beta^{RL}$ | $\alpha$ | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 8 |
| 19 | $\eta$ | - | $\beta_m$ ($\alpha_+$,$\alpha_-$,$\omega_1$,$\omega_2$fix) | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 5 |
| 20 | $\eta$ | $\omega^{RL}$ | $\beta^{RL}$ | $\alpha$ | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | $\gamma$ | 8 |
| 21 | $\eta$ | - | $\beta^{RL}$ | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 5 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 22 | $\eta$ | - | $\beta^{RL}$ | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | $\gamma$ | 6 |
| 23 | $\eta_{Gain}$,$\eta_{Loss}$ | - | $\beta^{RL}$ | - | $\omega^{WM}$ | $\beta^{WM}$ | $C_{WM}$ | - | 6 |
| 24 | $\eta$ | - | $\beta^{RL}$ | - | - | - | - | $\gamma$ | 3 |
| 25 | $\eta_{Gain}$,$\eta_{Loss}$ | - | $\beta^{RL}$ | - | - | - | - | - | 3 |
| 26 | $\eta$ | - | $\beta^{RL}$ | - | - | - | - | - | 2 |

## References


Alexander WH, Brown JW (2015) Hierarchical Error Representation: A Computational Model of Anterior Cingulate and Dorsolateral Prefrontal Cortex. Neural Comput 27:2354-2410.

Averbeck BB (2017) Amygdala and ventral striatum population codes implement multiple learning rates for reinforcement learning. IEEE Symposium Series on Computational Intelligence (SSCI):1-5.

Badre D, Doll BB, Long NM, Frank MJ (2012) Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron 73:595-607.

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. J Cogn Neurosci 28:333-349.

Balleine BW (2019) The Meaning of Behavior: Discriminating Reflex and Volition in the Brain. Neuron 104:47-62.

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.

Boorman ED, Behrens TE, Woolrich MW, Rushworth MF (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.

Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Blundell C, Hassabis D (2019) Reinforcement Learning, Fast and Slow. Trends Cogn Sci 23:408-422.

Caze RD, van der Meer MA (2013) Adaptive properties of differential learning rates for positive and negative outcomes. Biol Cybern 107:711-719.

Chelazzi L, Marini F, Pascucci D, Turatto M (2019) Getting rid of visual distractors: the why, when, how, and where. Curr Opin Psychol 29:135-147.

Collins AG, Frank MJ (2012) How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci 35:1024-1035.

Collins AG, Frank MJ (2013) Cognitive control over learning: creating, clustering, and generalizing task-set structure. Psychol Rev 120:190-229.

Collins AG, Brown JK, Gold JM, Waltz JA, Frank MJ (2014) Working memory contributions to reinforcement learning impairments in schizophrenia. J Neurosci 34:13747-13756.

Collins AGE, Ciullo B, Frank MJ, Badre D (2017) Working Memory Load Strengthens Reward Prediction Errors. J Neurosci 37:4332-4342.

Dajani DR, Uddin LQ (2015) Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. Trends Neurosci 38:571-578.

Donegan NH (1981) Priming-produced facilitation or diminution of responding to a Pavlovian unconditioned stimulus. J Exp Psychol Anim Behav Process 7:295-312.

Esber GR, Haselgrove M (2011) Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. Proc Biol Sci 278:2553-2561.

Failing M, Theeuwes J (2018) Selection history: How reward modulates selectivity of visual attention. Psychon Bull Rev 25:514-538.

Farashahi S, Rowe K, Aslami Z, Lee D, Soltani A (2017) Feature-based learning improves adaptability without compromising precision. Nat Commun 8:1768.

Frank MJ, Seeberger LC, O'Reilly R C (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science 306:1940-1943.

Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE (2007) Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A 104:16311-16316.

Gershman SJ, Daw ND (2017) Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. Annu Rev Psychol 68:101-128.

Hall G, Pearce JM (1979) Latent inhibition of a CS during CS-US pairings. J Exp Psychol Anim Behav Process 5:31-42.

Hassani SA, Oemisch M, Balcarras M, Westendorff S, Ardid S, van der Meer MA, Tiesinga P, Womelsdorf T (2017) A computational psychiatry approach identifies how alpha-2A noradrenergic agonist Guanfacine affects feature-based reinforcement learning in the macaque. Sci Rep 7:40606.

Kahnt T, Park SQ, Cohen MX, Beck A, Heinz A, Wrase J (2009) Dorsal striatal-midbrain connectivity in humans predicts how reinforce- ments are used to guide decisions. Journal of Cognitive Neuroscience 21:1332-1345.

Khamassi M, Enel P, Dominey PF, Procyk E (2013) Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. Prog Brain Res 202:441-464.

Khamassi M, Quilodran R, Enel P, Dominey PF, Procyk E (2015) Behavioral Regulation and the Modulation of Information Coding in the Lateral Prefrontal and Cingulate Cortex. Cereb Cortex 25:3197-3218.

Kruschke JK (2011) Models of attentional learning. In: Formal Approaches in Categorization (Pothos EM, Wills AJ, eds), pp 120-152: Cambridge University Press.

Lavie N, Fox E (2000) The role of perceptual load in negative priming. J Exp Psychol Hum Percept Perform 26:1038-1052.

Le Pelley ME, Pearson D, Griffiths O, Beesley T (2015) When goals conflict with values: counterproductive attentional and oculomotor capture by reward-related stimuli. J Exp Psychol Gen 144:158-171.

Lefebvre G, Lebreton M, Meyniel F, Bourgeois-Gironde S, Palminteri S (2017) Behavioural and neural characterization of optimistic reinforcement learning. Nature Human Behaviour 1:1-9.

Leong YC, Radulescu A, Daniel R, DeWoskin V, Niv Y (2017) Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. Neuron 93:451-463.

Namburi P, Beyeler A, Yorozu S, Calhoon GG, Halbert SA, Wichmann R, Holden SS, Mertens KL, Anahtar M, Felix-Ortiz AC, Wickersham IR, Gray JM, Tye KM (2015) A circuit mechanism for differentiating positive and negative associations. Nature 520:675-678.

Niv Y, Daniel R, Geana A, Gershman SJ, Leong YC, Radulescu A, Wilson RC (2015) Reinforcement learning in multidimensional environments relies on attention mechanisms. J Neurosci 35:8145-8157.

Noonan MP, Crittenden BM, Jensen O, Stokes MG (2018) Selective inhibition of distracting input. Behav Brain Res 355:36-47.

Oemisch M, Westendorff S, Azimi M, Hassani SA, Ardid S, Tiesinga P, Womelsdorf T (2019) Feature-specific prediction errors and surprise across macaque fronto-striatal circuits. Nat Commun 10:176.

Poldrack RA, Packard MG (2003) Competition among multiple memory systems: converging evidence from animal and human brain studies. Neuropsychologia 41:245-251.

Radulescu A, Daniel R, Niv Y (2016) The effects of aging on the interaction between reinforcement learning and attention. Psychol Aging 31:747-757.

Rmus M, McDougle SD, Collins AGE (2020) The Role of Executive Function in Shaping Reinforcement Learning. PsyArXiv:1-13.

Roelfsema PR, Holtmaat A (2018) Control of synaptic plasticity in deep cortical networks. Nat Rev Neurosci 19:166-180.

Rombouts JO, Bohte SM, Roelfsema PR (2015) How attention can create synaptic tags for the learning of working memories in sequential tasks. PLoS Comput Biol 11:e1004060.

Rusz D, Le Pelley M, Kompier MAJ, Mait L, Bijleveld E (2020) Reward-driven distraction: A meta-analysis. PsyArXiv https://doi.org/10.31234/osf.io/82csm.

Schonberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci 27:12860-12867.

Seo H, Cai X, Donahue CH, Lee D (2014) Neural correlates of strategic reasoning during competitive games. Science 346:340-343.

Sutton RS, Barto AG (2018) Reinforcement learning: An introduction., 2nd Edition: MIT Press.

Taswell CA, Costa VD, Murray EA, Averbeck BB (2018) Ventral striatum's role in learning from gains and losses. Proc Natl Acad Sci U S A 115:E12398-E12406.
Tomov MS, Truong VQ, Hundia RA, Gershman SJ (2020) Dissociable neural correlates of uncertainty underlie different exploration strategies. Nat Commun 11:2371.
van den Bos W, Cohen MX, Kahnt T, Crone EA (2012) Striatum-medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. Cereb Cortex 22:1247-1255.
van der Meer M, Kurth-Nelson Z, Redish AD (2012) Information processing in decision-making systems. Neuroscientist 18:342-359.
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. Front Behav Neurosci 9:225.
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. Behav Brain Res 355:76-89.
Voloh B, Watson MR, Koenig S, Womelsdorf T (2020) MAD saccade: statistically robust saccade threshold estimation via the median absolute deviation. Journal of Eye Movement Research 12.
Wang JX, Kurth-Nelson Z, Kumaran D, Tirumala D, Soyer H, Leibo JZ, Hassabis D, Botvinick M (2018) Prefrontal cortex as a meta-reinforcement learning system. Nat Neurosci 21:860-868.
Watson MR, Voloh B, Naghizadeh M, Womelsdorf T (2019a) Quaddles: A multidimensional 3-D object set with parametrically controlled and customizable features. Behav Res Methods 51:2522-2532.
Watson MR, Voloh B, Thomas C, Hasan A, Womelsdorf T (2019b) USE: An integrative suite for temporally-precise psychophysical experiments in virtual environments for human, nonhuman, and artificially intelligent agents. J Neurosci Methods 326:108374.
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. J Neurophysiol 116:796-811.
Wilson RC, Niv Y (2011) Inferring relevance in a changing world. Front Hum Neurosci 5:189.
Womelsdorf T, Thomas C, Parker J, Neuman A, König S, Banaie Boroujeni K, Hassani AS, Watson M, Hoffman KL (in preparation) A Cage-based Monkey Kiosk for Robust Cognitive Profiling and Enrichment
Worthy DA, Otto AR, Maddox WT (2012) Working-memory load and temporal myopia in dynamic decision making. J Exp Psychol Learn Mem Cogn 38:1640-1658.