Attentional Selection Can Be Predicted by Reinforcement Learning of Task-relevant Stimulus Features Weighted by Value-independent Stickiness

Matthew Balcarras[1*], Salva Ardid[1,2*], Daniel Kaping[1], Stefan Everling[3], and Thilo Womelsdorf[1,3]

Abstract

■ Attention includes processes that evaluate stimuli relevance, select the most relevant stimulus against less relevant stimuli, and bias choice behavior toward the selected information. It is not clear how these processes interact. Here, we captured these processes in a reinforcement learning framework applied to a feature-based attention task that required macaques to learn and update the value of stimulus features while ignoring nonrelevant sensory features, locations, and action plans. We found that value-based reinforcement learning mechanisms could account for feature-based attentional selection and choice behavior but required a value-independent stickiness selection process to explain selection errors while at asymptotic behavior. By comparing different reinforcement learning schemes, we found that trial-by-trial selections were best predicted by a model that only represents expected values for the task-relevant feature dimension, with nonrelevant stimulus features and action plans having only a marginal influence on covert selections. These findings show that attentional control subprocesses can be described by (1) the reinforcement learning of feature values within a restricted feature space that excludes irrelevant feature dimensions, (2) a stochastic selection process on feature-specific value representations, and (3) value-independent stickiness toward previous feature selections akin to perseveration in the motor domain. We speculate that these three mechanisms are implemented by distinct but interacting brain circuits and that the proposed formal account of feature-based stimulus selection will be important to understand how attentional subprocesses are implemented in primate brain networks. ■

INTRODUCTION

Selective attention can be defined as a set of processes that work around resource limitations by prioritizing processing to goal-relevant information (Womelsdorf & Everling, 2015; Tsotsos, 2011), while ensuring flexibility to adapt to new situations (Ardid & Wang, 2013; Kruschke & Hullinger, 2010; Dayan, Kakade, & Montague, 2000). Such a definition of attention implicitly assumes a continuous evaluation of the relevance of sensory information (Gottlieb, 2012; Kaping, Vinck, Hutchison, Everling, & Womelsdorf, 2011), which entails computing value predictions of stimulus features (Rangel & Clithero, 2014; Anderson, 2013; Chelazzi, Perlato, Santandrea, & Della Libera, 2013; Rushworth, Noonan, Boorman, Walton, & Behrens, 2011). Consistent with this suggestion, recent neurophysiological studies have shown that neural representations of stimulus value affect attentional search performance and gaze allocation in human participants (Anderson, Laurent, & Yantis, 2011; Tatler, Hayhoe, Land, & Ballard, 2011; Della Libera & Chelazzi, 2009) and underlie economic choices (Hare, Schultz, Camerer, O'Doherty, & Rangel, 2011; Padoa-Schioppa, 2011; Wunderlich, Rangel, & O'Doherty, 2010). Furthermore, neural correlates of those signals have been found in prefrontal and parietal neurons as well as in subcortical neural circuits (Cai & Padoa-Schioppa, 2014; Luk & Wallis, 2013; Peck, Lau, & Salzman, 2013; Kaping et al., 2011; Kennerley, Behrens, & Wallis, 2011; Peck, Jangraw, Suzuki, Efem, & Gottlieb, 2009). However, it is unclear how value-based learning relates to the covert attentional selection of stimulus features that precedes overt choices, as opposed to the learning of action values that immediately triggers overt choices (Glimcher, 2011; Lau & Glimcher, 2005). To elucidate the mechanisms that underlie attention, task paradigms and analyses need to isolate the learning of covert (attentional) stimulus selection from processes linked to overt choice such as perceptual discrimination and action planning (Rangel & Clithero, 2014).

In the decision-making domain, reinforcement learning (RL) provides a framework that links stimulus or action valuation to choice behavior (Rangel & Hare, 2010; Rushworth & Behrens, 2008). Commonly applied RL realizes goal-directed choices by (1) the continuous updating of value predictions of sensory features, (2) a softmax stochastic choice process among features that ensure performance accuracy while allowing for occasional

[1]York University, Toronto, Canada, [2]Boston University, [3]University of Western Ontario

*These authors contributed equally to this work.

Journal of Cognitive Neuroscience 28:2, pp. 333–349
doi:10.1162/jocn_a_00894

exploratory choices, and (3) rapid learning from the consequences (outcomes) of selections using prediction error signals (Rushworth & Behrens, 2008). These processing components could likewise account for the efficient top–down control of attention and may thus provide a framework to understand the interplay of attentional subprocesses (Womelsdorf & Everling, 2015; Wilson & Niv, 2011; Roelfsema & van Ooyen, 2005; Dayan et al., 2000). To test this hypothesis, we devised a feature-based reversal learning task for macaque monkeys that allowed quantifying whether commonly used RL frameworks help to understand how the learning of efficient attentional control is implemented and integrated during goal-directed behavior.

We found that the learning of attentional stimulus selections in nonhuman primates closely followed an RL model that acts on representations of a restricted set of task-relevant features, rather than on a representation of all stimulus and action items that could be linked to the decision outcome (Rangel & Clithero, 2014). However, we also show that a feature-based RL model of attention needs to be supplemented with a value-independent stickiness process to account for non-randomly distributed errors during asymptotic behavior.

METHODS

Experiments were performed in two male macaque monkeys following guidelines of the Canadian Council of Animal Care policy on the use of laboratory animals and of the University of Western Ontario Council on Animal Care. Monkeys sat in a custom-made primate chair viewing visual stimuli on a computer monitor (85-Hz refresh rate, distance of 58 cm) in a sound-attenuating isolation chamber (Crist Instrument Co., Inc., Hagerstown, MD). The monitor covered 36° × 27° of visual angle at a resolution of 28.5 pixel/deg. Eye positions were monitored using a video-based eye-tracking system (ISCAN, Woburn, MA; sampling rate: 120 Hz) and were calibrated before each experiment to a 5-point fixation pattern. During the experiments, eye fixation on a 0.2° gray square was controlled within a 1.4–2.5° radius window. Monitoring of eye positions, stimulus presentation, and reward delivery were controlled through MonkeyLogic (open-source software, www.monkeylogic.net) running on a PC Pentium III (Asaad & Eskandar, 2008). Liquid reward was delivered by a custom-made, air compression-controlled, mechanical valve system with a noise level during valve openings of ≤17 dB within the isolation chamber.

Task Design

We trained the monkeys on a feature-based reversal learning task (Figure 1A). The task required monkeys to fixate and covertly attend to one of two peripherally presented stimuli. Stimuli had different colors, and only one color was associated with reward across trials within a block. To obtain reward, the animals had to discriminate a transient rotation of the attended stimulus. Rotations also occurred in the stimulus with the non-reward-associated

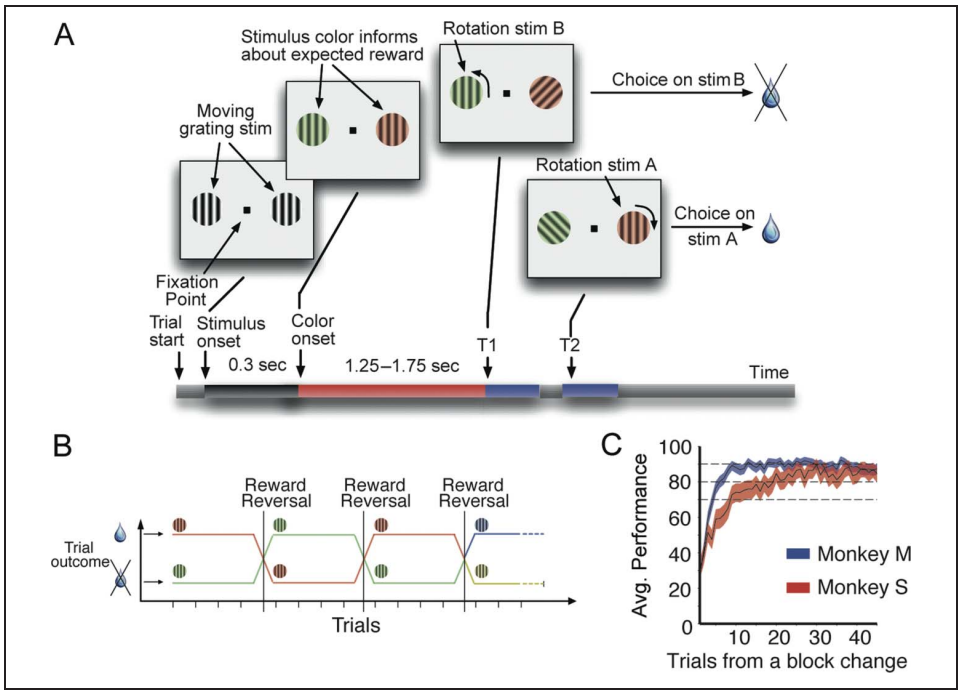


Figure 1. Feature-based attentional learning task. (A) Uncued task design. Monkeys learned by practice that only the color dimension of the stimuli was associated with reward, whereas other features (location, rotation direction, or time onset of the rotation) were completely irrelevant. A proper allocation of covert attention allowed monkeys to successfully discriminate a transient rotation in the relevant stimulus while ignoring that of the distractor. Monkeys reported their response with an upward versus downward saccade according to the rotation direction, which was reversed in the two monkeys. (B) Color–reward associations were changed in blocks of trials. (C) Average performance for Monkeys M and S as a function of trial number in the block. The shaded area denotes the 95% confidence interval. Avg. = average; stim = stimulus.

color. Monkeys indicated their choice by making a saccadic eye movement to one of two response targets presented 6.7° above or below the fixation point (clockwise/counterclockwise rotations were mapped onto upward/downward saccades for one monkey and onto downward/upward saccades for the second monkey). In each block of trials, reward was associated only with one color. No reward was given to rotation discriminations of the stimulus with the nonrewarded color. Rotation direction (clockwise vs. counterclockwise), location (right vs. left), and the time onset of rotation of the stimulus with rewarded and nonrewarded color (first vs. second vs. simultaneous) changed randomly across trials. In each trial, the stimulus with the rewarded color and the stimulus with the nonrewarded color rotated in opposite directions.

The event sequence in a trial was as follows (Figure 1A). Monkeys initiated trials by directing and maintaining their gaze on a centrally presented, gray fixation point (on a black, 0.6-cd background), followed 0.3 sec later by the onset of two stimuli. Within the stimulus aperture, motion direction of a grating to the left from fixation was always to the upper left (−45° from vertically up), and motion direction of the stimulus on the right side from fixation was always to the upper right (+45° from vertically up). After 0.4 sec, the stimuli were colored. The rotation of the rewarded and nonrewarded stimulus occurred either at 0.75 sec or at 1.35 sec. Trials in which the stimulus with the rewarded color rotated before or after the stimulus with the nonrewarded color were counterbalanced. In 10–50% (on average, 30%) of all trials, the rotation of the stimuli with the rewarded and nonrewarded color occurred at the same time (1 sec after the color onset). Trials with rotations at the same time were introduced to validate that animals succeeded to select the relevant stimulus before discriminating the relevant rotation direction. This manipulation ensured that monkeys could perform the task only above chance when the stimulus with the rewarded color (and not the distractor, or a response direction, or other aspects) was selected. After stimulus rotation, animals made a saccadic response toward either of two target dots located vertically, above versus below, with respect to the fixation point, to report the rotation direction of the chosen stimulus. To obtain reward, a saccade had to be made 0.05–0.5 sec after rotation onset of the stimulus associated with the rewarded color. Animals received a fluid reward with a delay of 0.4 sec after the saccadic response.

Within an experimental session, the color–reward association was alternated in blocks of 60–100 trials, either maintaining the same pair of colors or by introduction of a new pair (Figure 1A). After a minimum of 60 trials, a new block was introduced as soon as either of three performance criteria was achieved: (1) running average performance (over 15 preceding trials) of rewarded correct sensory–response associations relative to unrewarded incorrect choices exceeded 80%, (2) a total number of 60 rewarded trials, or (3) a total number of 100 trials independent on whether the choice was rewarded. Each experimental session also included shorter blocks of (n = 30) cued trials, which, besides the cue instruction, had identical timing and stimulus events as the uncued trials described above. In cued trials, the fixation point was colored to match the color of one of the peripheral stimuli, which was indicative of that stimulus being relevant. Stimulus colors used in the cued trials were never used in the uncued trials. Cued trials were not analyzed in this report.

Stimuli

We used square wave gratings with rounded-off edges for the peripheral stimuli (Figure 1A), moving within a circular aperture at 1 deg/sec, a spatial frequency of 1.4 Hz/deg, and a radius of 2.2°. Gratings were presented at 4–6° eccentricity to the left and right of fixation. The grating on the left (right) side always moved within the aperture upwards at −45° (+45°) relative to vertical. The angle of rotation ranged between ±13° and ±19°. The rotation proceeded smoothly from the standard direction of motion toward maximum tilt within 60 msec, staying at maximum tilt for 235 msec, rotated back to the standard direction within 60 msec, and continued moving at their prechanged direction of motion at −45° or +45° relative to vertical thereafter.

Analysis of Performance and Learning within Blocks

Data analysis was done with custom written MATLAB scripts (The MathWorks, Inc., Natick, MA). Analysis was performed on n = 200 experimental sessions (n = 100 sessions for Monkey M and n = 100 sessions from Monkey S). To identify at which trial during a block the monkeys showed statistically reliable learning, we analyzed the monkeys' trial-by-trial choice dynamics using the state space framework introduced by Smith and Brown (2003) and implemented by Smith et al. (2004). This framework entails a state equation that describes the internal learning process as hidden Markov or latent process and is updated each trial. The learning state process estimates the probability of a correct response in each trial and thus provides the learning curve of subjects (see, e.g., Wirth et al., 2003). The algorithm estimated learning from the perspective of an ideal observer that takes into account all trial outcomes of the subjects in a block of trials to estimate the probability that the outcome in a single trial is correct or incorrect based on the modeled learning state process. This probability is then used to calculate the confidence range of observing a correct response. We defined the learning trial as the earliest trial during a block at which the lower confidence bound of the probability for a correct response exceeded the chance level (here: p = .5).

More specifically, the algorithm defines the learning state process as a random walk whereby each trial's probability of a correct response depends on the previous trials probability or on the chance level in case there was no previous trial's probability at the beginning of blocks. According to this formulation, the subjects' choices across trials follow a random strategy. The mean of the random process reflects the current probability for a correct response. The variance of the random process determines how fast the learning state process can change from trial to trial and thus how rapidly learning can take place (see Smith et al., 2004). The expectation maximization algorithm is used to estimate the mean and variance of the random process by maximum likelihood (Dempster, Laird, & Rubin, 1977) to derive the probability to observe a correct response in each trial as a function of the trial number (Smith & Brown, 2003). A forward filter estimates the variance and mean of the value of the Gaussian random variable from the first trial to the current trial. This forward process reflects a state estimate from the perspective of the subject performing the task. An additional smoothing algorithm takes the perspective of an ideal observer and estimates the current trials' mean and variance of the state process using data from all trials. The estimates of both, the forward filter and the smoothing process, are then used to calculate the probability density for the correct response probability at each trial (please see Smith et al., 2004, Equations 2.1–2.4 for details). The aforementioned procedure provides the learning curve, that is, it provides for each trial the probability of a correct response given the sequence of correct and incorrect choices of the monkey. To identify the first trial in a block at which an ideal observer knows with $p \geq .95$ confidence that learning has taken place, we calculated the lower confidence bound and identified the first trial where it exceeded the chance performance as the learning trial ($p = .5$), the first "IO95" learning trial (see Smith et al., 2004). This corresponds to a .95 confidence level for an ideal observer to identify learning.

Logistic Regression Analysis

We developed a logistic regression analysis (using the glmfit function of MATLAB, The MathWorks, Inc.) over the complete set of trials under consideration to check whether RL mechanisms were overall consistent with monkeys' performance in the task and, if so, to infer specific RL characteristics that we could then use in the implementation of RL models (see below).

In particular, we analyzed and ranked the predictive power for attentional selection of stimulus features. We tested four different versions of the logistic regression analysis depending on how features in trial T predicted attentional selection of one of the two stimuli, and the choice, in the following trial $T + 1$: Version 1 features predict attentional selection of the stimulus they belong (in trial $T + 1$) if they formed part of the previously selected stimulus (inferred from choice in trial T) regardless of outcome (note that this is a control case in which the regression analysis is actually not consistent with RL mechanisms); Version 2 features predict attentional selection of the stimulus they belong (in trial $T + 1$) if those features formed part of the previously selected stimulus (in trial T) and that trial was rewarded (this case is compatible with RL, such that positively correlated feature–reward associations are reinforced for subsequent attentional selection); Version 3 features predict attentional selection of the stimulus they belong (in trial $T + 1$) if those features did not form part of the previously selected stimulus (in trial T) and that trial was not rewarded; and Version 4 that combines the previous two conditions, so features predict attentional selection of the stimulus they belong (in trial $T + 1$) if they formed part of the previously selected stimulus (in trial T) and that trial was rewarded as well as if they did not form part of the previously selected stimulus and the trial was not rewarded.

Interestingly, results from this latter condition were best predictive of monkey choices (Figure 2, other conditions not shown), which suggest a value-update generalization of the features in the two stimuli, even when monkeys in each trial only acted on one of the two stimuli.

We performed the ranking on half of the sessions (odd session numbers) and validated this ranking of features on the other half of the sessions (even session numbers). Figure 2 shows the proportion of choices explained with respect to chance level from a collection of regression analyses, in which each analysis included one more regressor than the previous, beginning with the regressor with the largest predictive power and continuing according to the predictive power ranking, until all regressors were taken into account. These results confirmed that colors were the best predictors of next choices, supporting the hypothesis of value-based covert attentional selection guiding monkeys' behavior.

Because of the two-alternative choice, the chance level was computed as 50% of the trials in which at least one of the features in the feature set was present in trial $T + 1$ and formed part of the stimulus associated with reward by task design in trial T (hence predictive of choice in trial $T + 1$; also note that this did not necessarily imply that monkeys acted on the stimulus associated with reward in trial T). For instance, for Monkey M, the first regressor (color C3) correctly predicted 21.35% of the next choices. This represented 86.44% of the trials in which the color C3 determined the stimulus associated with reward by task design (24.7% of the whole set of trials, far beyond the chance level at 12.35%). Importantly, the proportion of trials explained initially grew at a similar pace while including colors in the regression analysis but then drastically stopped, showing that including other features did not improve further the predictive power of monkey choices (Figure 2). Note also the increasing separation

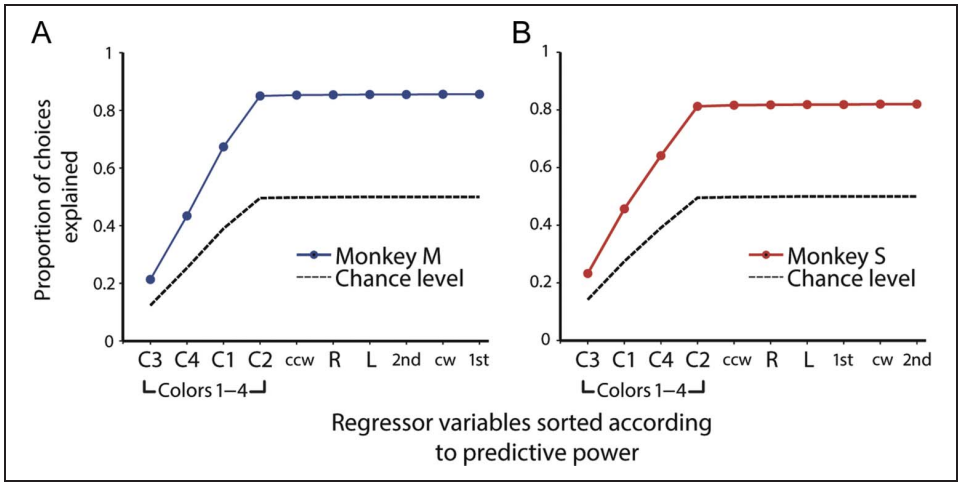


Figure 2. Logistic regression analysis of monkey performance in the task. We ranked the predictive power of each stimulus feature for subsequent performance according to a logistic regression analysis (see Methods for details). In both monkeys, the regression that included only color information was the best predictor of next choices, which confirmed that nonhuman primates primarily utilize reward-associated stimulus features to guide their covert attentional selection. (A) Monkey M. (B) Monkey S. In both panels, labels "C1," "C2," "C3," and "C4" in the figure denote the individual stimulus colors used; "ccw" and "cw" denote stimulus rotation direction; "R" and "L" denote stimulus locations; and "1st" and "2nd" denote the relative time of movement onset of the rewarded stimulus in relation to that of the distractor.

with respect to the chance level when incorporating color features to the regression analysis, until it reached its maximum when all colors were included in the analysis. This separation remained the same after including non-color features to the regression analysis (Figure 2).

RL Modeling

To model monkeys' behavior and the processes related to covert attentional selection, we used Rescorla–Wagner type of RL employing standard Q-Learning and Boltzman softmax selection algorithms (Glimcher, 2011). We initially compared two distinct value-based RL models that differed in whether a restricted, optimal feature set (composed only of stimulus colors) was used to select one of the two peripheral stimuli ("feature-based RL") or whether the selection process considered all features ("nonselective RL").

To explain a specific pattern of error trials shown by the monkeys, which was not reproduced by value-based models, we explored additional non-value-based mechanisms: First, we accounted for an influence of selection perseveration that is unaffected by values, which has been shown to improve action selection (Lau & Glimcher, 2005). This value history model (Figure 6A) transforms feature values into probabilities of attentional selection as the feature-based RL does, but it then incorporates a weighted bias toward whatever feature was selected on the previous trial.

The second extension of feature-based RL, the hierarchical value-history model, is similar to the previous value history model, but in this formulation, the value-based selection process is concatenated with a subsequent final attentional selection between the selected feature in the previous trial versus the current trials' value-based selected feature (Figure 6B). This sequential selection can therefore be conceived of as a hierarchical two-step decision process.

Third, we quantified the influence of a mechanism that dynamically adjusted the exploration versus exploitation trade-off based on performance. This adaptive selection model incorporated a meta-learning parameter that scaled up or down the nonlinearity in the transformation from value to probability of attentional selection according to reward outcome (Figure 6C). Thus, when model performance is low, typically at the beginning of a block, more exploratory behavior is produced because of a low β value, because it increases the stochasticity of selection among features. As rewarded outcomes become more frequent, β increases, which makes attentional selection be more deterministic.

In a fourth model extension, we incorporated non-value-based noise into the attentional selection process (Figure 6D). In this intrinsic noise model, such noise is evenly distributed among all stimulus features. Thus, there is no dependence on value, reward, or selection history in this module of the model, but rather an explicit influence of noise, intrinsic to the transformation of value-based selections to motor commands because of influences, such as decreased motivation, imperfect sensory-motor mappings, or selection biases, among others, under the assumption that these influences do not show a preference for specific features in the internal model representation of the task.

RL Model Algorithms

In its basic form, the value of any predictor of reward (Q_i) is updated on the next time step (trial) from its previous value through the scaled reward-prediction error: the difference between the binary reward outcome (R, either 0

or 1) and the predictor itself. The scaling factor (α) represents the learning rate:

$$Q_i(n+1) = Q_i(n) + \alpha[R(n) - Q_i(n)] \quad (1)$$

We implemented RL models that assumed value generalization. Thus, all stimulus features associated with the selected stimulus updated their value according to Equation 1. Stimulus features associated with the other stimulus were updated according to

$$Q_i(n+1) = Q_i(n) + \alpha[1 - R(n) - Q_i(n)] \quad (2)$$

Our RL approach assumed that performance in a trial only depended on a correct covert attentional selection of the relevant stimulus, which implied an infallible rotation discrimination and its associated saccadic response.

The "feature-based RL" took into account only the systematically relevant color dimension as predictor of attentional selection and, therefore, of reward (Figure 3A). In contrast, in the "nonselective RL," all stimulus features (colors, locations, rotation directions, and time onsets of the rotation) were considered potential predictors of reward (Figure 3B).

The final attentional selection of one stimulus against the other obeyed a covert, value-based softmax decision-making process acting on the feature space, in particular, on nonlinearly transformed values that represented the probabilities of selecting different stimulus features, according to the Boltzmann equation:

$$P_i(n) = \frac{e^{\beta Q_i(n)}}{\sum_j e^{\beta Q_j(n)}} \quad (3)$$

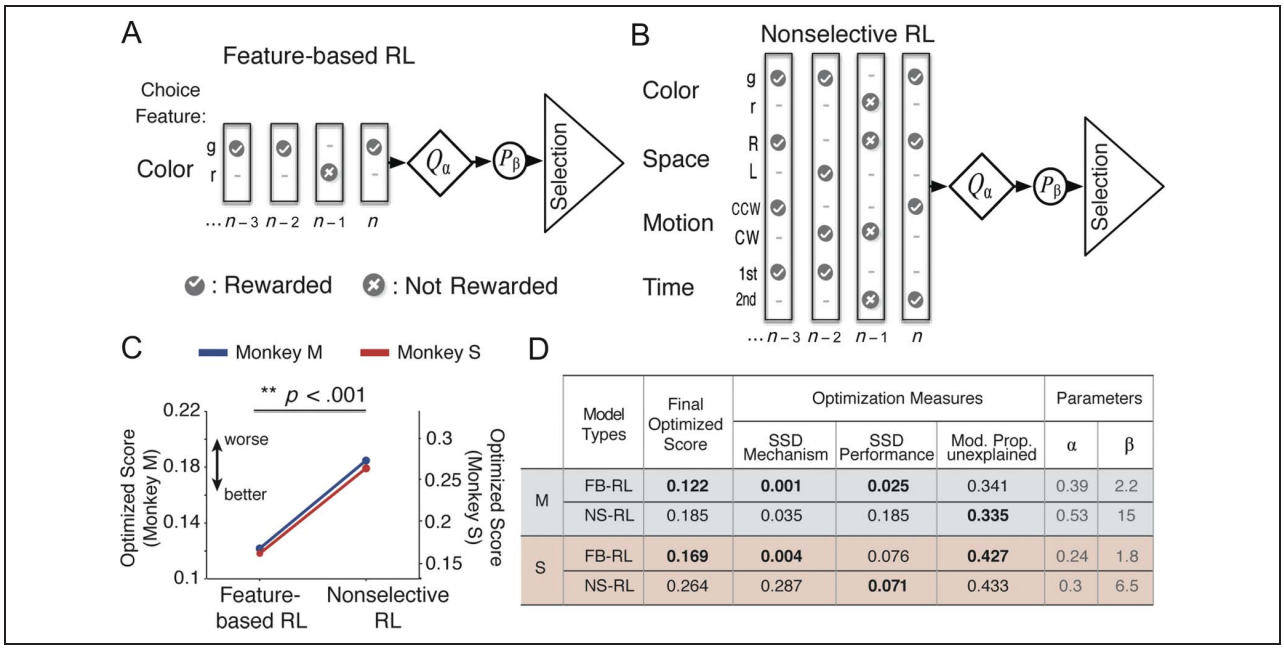
where β represents the inverse temperature and establishes the strength of the nonlinearity. The two RL models thus included two free parameters (α and β) that we optimized to best predict monkey behavior on a trial-by-trial basis (Figure 3).

Value History Model

The first extension of the feature-based formulation introduces an explicit factor that influences the value-based selection mechanism by biasing the selection toward the feature that was selected previously, irrespective of whether it was rewarded (Figure 6A). The selection of this value history model is formally implemented as:

$$P_i(n) = \frac{e^{\beta Q_i(n)} + e^{\gamma \delta_{ik}} - 1}{\sum_j e^{\beta Q_j(n)} + e^{\gamma \delta_{jk}} - 1} \quad (4)$$

where, in the γ term, k represents the previously selected feature and appears inside a Kronecker delta function,



	Model Types	Final Optimized Score	Optimization Measures			Parameters	
			SSD Mechanism	SSD Performance	Mod. Prop. unexplained	α	β
M	FB-RL	**0.122**	**0.001**	**0.025**	0.341	0.39	2.2
	NS-RL	0.185	0.035	0.185	**0.335**	0.53	15
S	FB-RL	**0.169**	**0.004**	0.076	**0.427**	0.24	1.8
	NS-RL	0.264	0.287	**0.071**	0.433	0.3	6.5

Figure 3. RL model schemes and results. (A) Feature-based RL tracks across trials' (…, $n - 1$, n) reward-dependent values (Q) only for the relevant stimulus features (colors, in the example: g = green, r = red) and nonlinearly transforms them into choice probabilities (P) of attentional selection through a softmax function. (B) Nonselective RL works the same but tracks values for all task features, denoted as Color, Space (stimulus location), Motion (rotation direction), and Time (order of each stimulus rotation onset). (C) The performance of the optimized feature-based RL is better than the performance of the optimized nonselective RL for both monkeys (left, x axis: Monkey M; right, x axis: Monkey S). (D) Optimization scores and parameters for feature-based (FB) RL and nonselective (NS) RL for Monkey M (blue shaded) and Monkey S (red shaded). Multiple scores were used to explicitly account for different aspects of monkeys' behavior and to directly test the predictive power of each RL mechanism (see text and Methods for details). Lower scores denote better model prediction. **Bold font** highlights best scores relative to the alternative model. SSD denotes normalized SSD in [0,1].

which takes a value of 1 if i is equal to k or 0 otherwise. The term -1 is included to remove any impact of the γ term when γ is 0. The effect of γ can be described as an increase in the probability to reselect the immediate previous selection, which in principle might be beneficial to diminish the impact of noise in the value system implementation, at the cost of a reduced celerity in the adaptation to changed feature–reward contingencies.

Hierarchical Value-History Model

As indicated above, the second extension of the model-based RL is similar to the value history model, but in this formulation, the selection process based on values is concatenated with a subsequent selection between the feature choice of the previous trial and the current trial value-based selected feature (Figure 6B). This sequential selection can be conceived of as a hierarchical two-step decision process. The first process fully corresponds to the feature-based selection process defined in Equation 3. From this selection, feature k is selected with "confidence" P_k dictated by the softmax function and used in a second step to compete with the previously selected feature P_l (if feature l is different than feature k):

$$P'_k = \frac{P_k}{P_k + P_l} \tag{5}$$

versus

$$P'_l = \frac{P_l}{P_k + P_l} \tag{6}$$

where $P_l = e^{\gamma} - 1$. When the value-based selected feature k and the previously selected feature l are the same, both terms add together, and the probability to select the feature trivially collapses to 1.

Adaptive Selection Model

The third extension of the feature-based formulation introduces a mechanism that adjusts the probabilistic nature of the value-based selection process to either trigger more exploratory selections or to more deterministically follow the valuation mechanism (corresponding to an exploitation regime with high confidence; see Figure 6C). The selection of values in the adaptive selection model uses Equation 3, but with the difference that β is not a constant but instead obeys an equation similar to the Q values (Equations 1 and 2; note that R is a binary teaching signal and then only one of the two terms in Equation 7 is different than zero in each trial):

$$\begin{aligned}\beta(n+1) = \beta(n) &+ R(n)\mu[\beta_H - \beta(n)] \\ &- [1 - R(n)]\mu\beta(n)\end{aligned} \tag{7}$$

where μ is the rate of change of β. β values are bounded between 0 and β_H. β tends to either one or the other depending on the outcome (R). If the outcome is 1, then β grows toward β_H; otherwise, it decreases to 0. Thus, after positive outcomes, the impact of β is to make the softmax function (P_i above) more similar to a winner-take-all but to otherwise encourage more exploratory behavior. Therefore, this model becomes more or less confident on the value system depending on outcome evaluation.

Intrinsic Noise Model

The fourth extension assumes that part of behavioral variability is in principle not explainable by value-based updating and selection mechanisms but rather is because of random behavioral variability and evenly distributed among features (Figure 6D):

$$P_i(n) = \frac{P_R}{N_F} + (1 - P_R)\frac{e^{\beta Q_i(n)} - 1}{\sum_j e^{\beta Q_j(n)} - 1} \tag{8}$$

The term P_R denotes the random behavioral probability that is evenly distributed among task features (N_F refers to the number of those). Note that -1 is introduced to remove the contribution of Q-values that are equal to 0. The value system is scaled down by the factor $1 - P_R$. This random weighting factor could theoretically fit the data better compared with the pure value-based model if the noise significantly splits into two parts: one noise component in the softmax (among Q values that are not strictly 0) and another noise component that is non-value-based. This is because a single β parameter in principle does not necessarily capture the two sources of noise at once but instead is designed to capture value-based stochasticity. This intrinsic noise model is similar to the value history model by adding a non-value-based process that competes with value, but for the intrinsic noise model, the non-value-based process operates at random among features instead of favoring the previous attentional selection.

Model Evaluation and Optimization

Three independent criteria (outlined in detail below) were combined to evaluate RL models. Such a multiscore evaluation was critical to (1) account for the dynamics of learning of monkeys in the task (performance sum of squared differences [SSD]), (2) analyze the plausibility of an RL mechanism for explaining monkey performance (Mechanism SSD), and (3) maximize the total number of trials in which monkeys and model performance matched, corrected to penalize model biases against the least frequent outcome (i.e., the overall proportion of trials explained was corrected by subtracting the highest between the proportion of false positives and false negatives).

The first score represented the SSD between the block-averaged performance of the model with respect to the monkey over the same blocks of trials (Figures 4A and

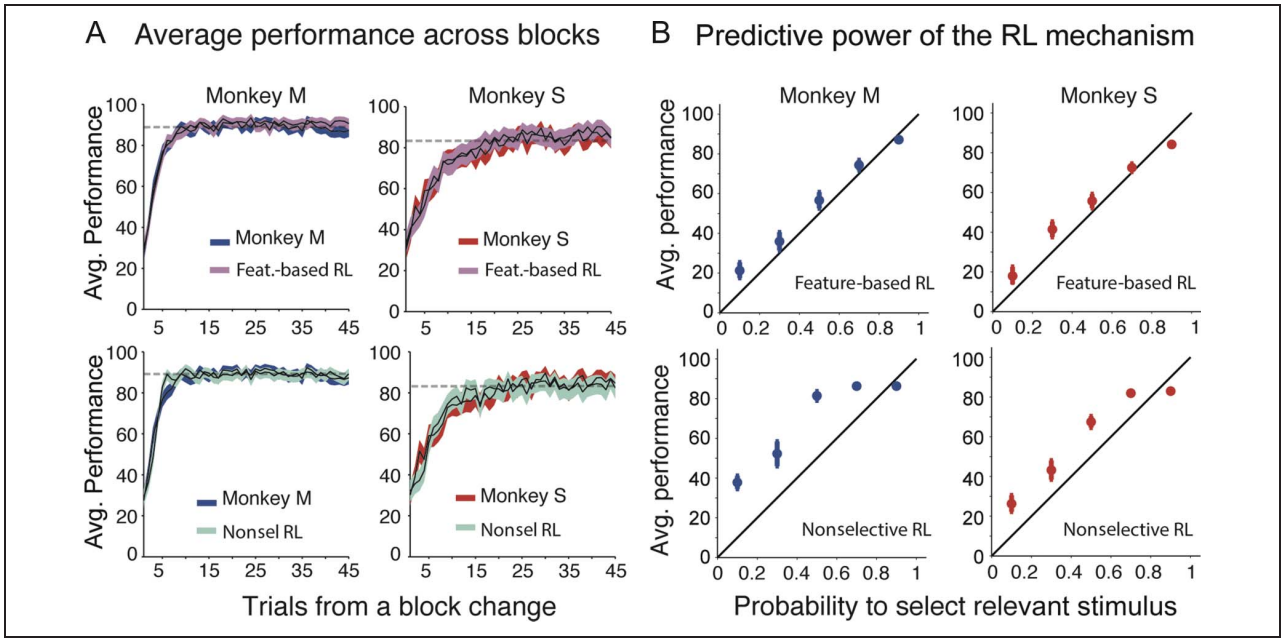


Figure 4. Performance of feature-based versus nonselective RL systems. (A) Average performance and its 95% CI (shaded area) as a function of trial order within a block predicted from feature-based RL (top row) and nonselective RL (bottom row) for Monkeys M (left) and S (right). The normalized SSD between the performance of the monkey and models served as one model evaluation criterion, labeled SSD Performance in Figure 3D. (B) Monkeys' averaged performance and its 95% CI (error bars) against the likelihood to select the relevant stimulus according to models. The panels show the performance of the monkeys (y axis) corresponding to five bins that fully span the range of the probability to select the relevant stimulus. A plausible model candidate requires the model's likelihood and monkey's performance to match each other. The degree to what this happens is quantified by the normalized SSD (labeled SSD Mechanism in Figure 3D). The panels are arranged as in A: feature-based RL (top row), nonselective RL (bottom row), Monkey M (left column), and Monkey S (right column).

7A and C). The second score quantified the extent to which the RL mechanism employed by a model was compatible with monkeys' behavior (Figures 4B and 7B and D). The average model performance only depends on the probability to select the relevant stimulus, and a direct test of this mechanism can be applied to monkeys' behavior: We binned the probability to select the relevant stimulus and computed for them the averaged monkeys' performance as well as its 95% confidence intervals (CIs). If the averaged performance of the monkeys was largely different from the probability to select the relevant stimulus according to the model, we can then conclude that such mechanism would not be fully compatible with monkeys' behavior, and this is visualized by deviations from the diagonal in Figures 4B and 7B and D. After calculating our measure of Performance SSD and Mechanism SSD, we normalized these scores across all models and parameter sets (independently in each monkey) to ensure that all scores were bounded in the same range [0,1].

The third measure evaluating the model performance compared the outcome experienced by the monkey on every trial with that of the model and calculated the total proportion of correctly matching trials. The common denotation of this measure is proportion of total explained trials. We modified this score to correct for the fact that it is important for a model to not only predict a high proportion of trials correctly but also ideally predict the correct proportion of rewarded and unrewarded trials, avoiding any potential bias. For example, a toy model that merely predicts a rewarded choice on every trial would provide no insight into the mechanisms driving monkey behavior but would report a total proportion explained > 80% because of the overall high proportion of rewarded behavior shown by the monkeys. We then corrected the total explained score by the proportion of false positives or false negatives (whichever was higher) to provide a single score that combines both raw explanatory power and a measure of predictive accuracy. The score appeared inverted (i.e., 1-score, corrected proportion of unexplained trials, so a lower score reflected a better model performance) to be in agreement with the two previously described scores.

To summarize, we used the three model performance scores to capture three distinct characteristics of the monkeys' behavior that standard model comparison methods (e.g., Bayesian Information Criterion/Akaike Information Criterion) would, at least partially, fail to capture (see below). First, the performance score accounts for the dynamics (trial-by-trial variation) in average performance in the course of a block. Second, the scoring explicitly tests the actual RL mechanism, that is, how well the probability to select the relevant stimulus then

translates in the average performance shown by the monkeys, under the model assumption that, if a model is "correct," it should reflect a 1:1 mapping. Finally, the "corrected total explained" performance score accounts for the same overall average performance in the task, while correcting for the different proportions of correct and error trials.

Models were fit to the behavioral data of the monkeys by performing a grid search across a broad region of the parameter space, avoiding exploring regions not sensitive to be feasible, such as those corresponding to extremely large values of the inverse temperature parameter. We first explored this parameter region for a given resolution (e.g., in 0.1 steps for the learning rate) and then zoomed in the area of interest, which was around the best scoring value that we found, using finer resolutions until reaching the point of no further improvement (e.g., 0.01 steps for the learning rate). Grid searches following this procedure produce analogous results to other methods, such as gradient descent (e.g., Donoso, Collins, & Koechlin, 2014; Collins & Frank, 2013). We validated the results of our optimization procedure through cross-validation between odd and even numbered sessions. Model performance was first assessed using odd numbered sessions (calibration data set) of monkey data by calculating the mean score for each parameter set across the three different measures, with each score representing the mean of 10 model replications to diminish the impact of fluctuations because of the stochasticity in the model. Please note that the results were not different when using a Bootstrap procedure with $n = 100$ replications. Then, best aggregate scores for each model computed on odd numbered sessions were used to assess model performance on even numbered sessions (test data set). Cross-validation of scores confirmed that parameters were not fit to nonsystematic behavior (e.g., which would have followed from overfitting) but instead represented a generalizable version of the model (Ahn, Busemeyer, Wagenmakers, & Stout, 2008).

Analysis of Error Patterns

Consecutive unrewarded trials during asymptotic performance (toward the end of a block, after the learning period; Figure 4A) were unlikely events in feature-based and nonselective RL systems (Figure 5), because feature values were very dissimilar and changed only minimally at asymptote, so errors were only because of the stochasticity of the selection process under such conditions. This suggests a random and independent distribution of errors during this period, which would be expected to happen also in monkey behavior if it would follow directives of the feature-based or nonselective RL system.

To test this null hypothesis, we counted all errors made during asymptotic behavior in a blockwise fashion and calculated the proportion of errors occurring in sequences of increasing length. To compare with a random distribution, we subtracted the proportion of errors for each error sequence from the theoretical proportion given by a random distribution. This transformation eased the identification of clusters of errors (i.e., unrewarded trials made consecutively), which occurred in monkeys more frequently than predicted by the stochasticity of learned values according to RL models (Figure 5). This finding suggested an additional selection mechanism influenced by non-value-based sources (Figure 6).

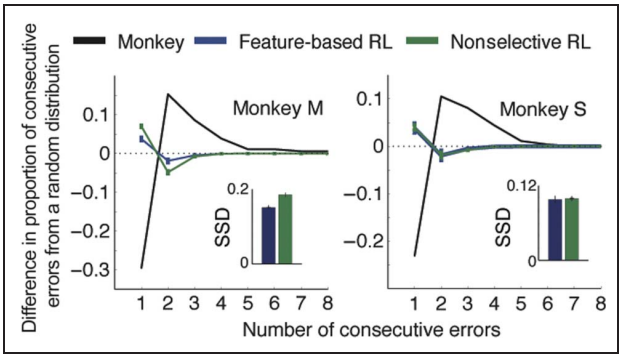


Figure 5. Failure of feature-based and nonselective RL systems to account for the pattern of consecutive errors shown by the animals during periods of asymptotic performance. The panel shows how the proportion of consecutive errors (*x* axis) by monkeys (M: left, S: right) deviated from what would be expected if errors were generated by a random process (dotted line). Feature-based RL (in blue) and nonselective RL (in green) failed to capture this error pattern. The inset bar panels show the SSD between the error pattern of monkeys and models. Errors represent *SEM*.

RESULTS

We devised a reversal learning task for macaques that isolates the covert attentional selection of relevant sensory (color) information from the perceptual discrimination (rotations) and action planning (saccades) processes directly involved in overt decision-making (Figure 1A). Covert attention was required to select one of two peripherally (left and right) presented stimuli for prioritized processing. Overt decision-making was required to obtain reward through discriminating a transient (clockwise/counterclockwise) rotation of the stimuli by making an (upward/downward) saccadic eye movement. Monkeys were rewarded only if the decision about the rotation was performed on one of the two stimuli with no reward given if the animal acted on the alternative stimulus. The rewarded stimulus was defined by its color with the reward-associated color changing between blocks of trials. The task design ensured that the stimulus color varied independently from (1) the stimulus location (right or left), (2) the decision variable of the overt choice (clockwise or counterclockwise rotation) that eventually provided the outcome, (3) the action plan (upward or

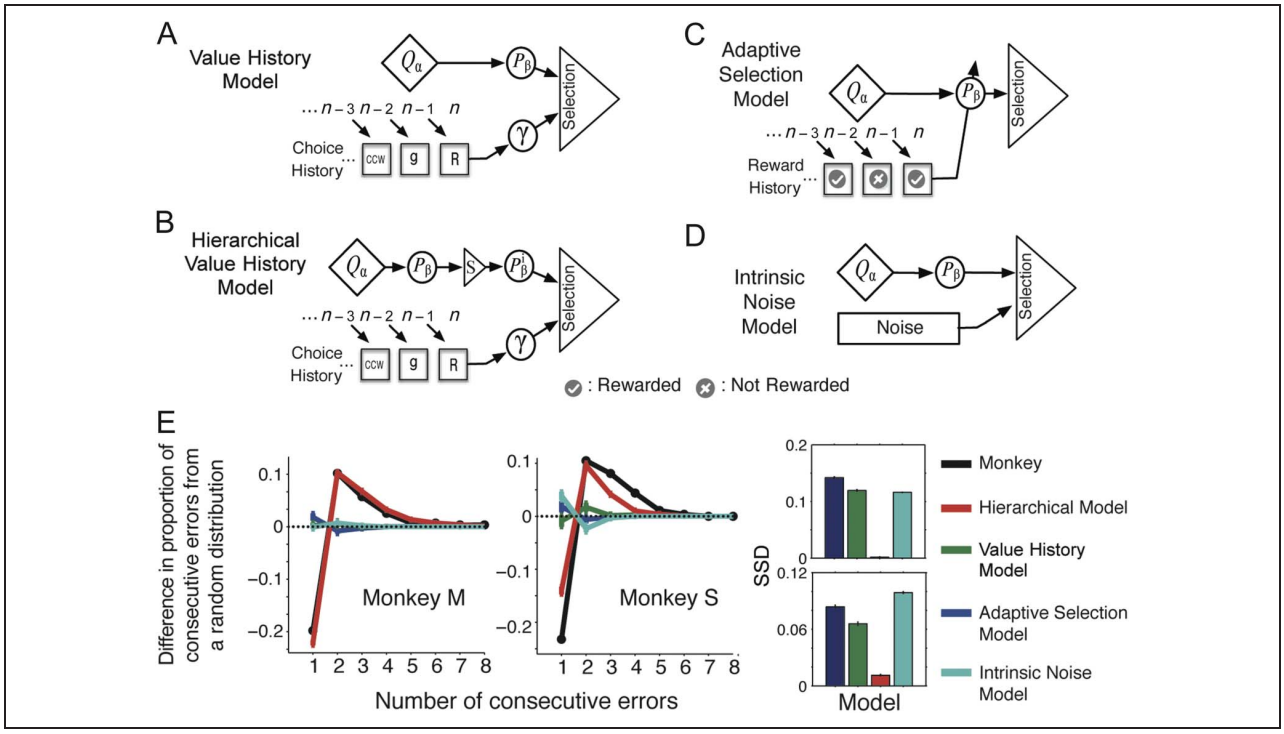


Figure 6. Schemes of extended feature-based RL systems and results from the analysis of consecutive errors in them. (A) The value history model incorporates in the covert attentional selection an influence toward previous choice. (B) The hierarchical value-history model first makes a value-based selection (P_i) and then weights it by the previous selection in a second selection step. (C) For the adaptive selection model, the transformation of Q-values to choice probabilities is dynamically shifted by the reward history. (D) The intrinsic noise model assumes that part of monkeys' performance stochasticity is independent of the value-based influences and distributes evenly among stimulus features. (E) Distribution of consecutive errors for Monkeys M and S, for each of the extended models shown in A–D. The two left panels show how the proportion of consecutive errors by monkeys (M: left, S: center) deviated from the proportion of errors that would be expected if errors were generated by a random process (as in Figure 5). Only the hierarchical model captured this property of animals' behavior. The bar plots (right) quantify the SSD between the error pattern of monkeys (M: top, S: bottom) and models. Errors represent *SEM*.

downward saccades) used to indicate overt choice, and (4) the three possible time points at which the stimulus rotation could occur. Thus, in contrast to previous learning paradigms in nonhuman primates (Lau & Glimcher, 2008; Sugrue, Corrado, & Newsome, 2004), the actual reward associated feature (color) was independent of action, location, and timing. The task enforced learning reward predictions about specific colors by changing the reward-associated color after a performance criterion, or a maximum number of trials was reached in a block of trials with constant color–reward association (see Figure 1B and Methods).

Monkeys Successfully Use Feature Values to Guide Attention

Both monkeys were successful in 82.5% of trials (Monkey M = 84.7% of 84,417 trials; Monkey S = 80.3% of 86,689 trials). Within blocks, monkeys required, on average, 12.5 trials to reach a performance level of 80% rewarded trials when measuring average performance using a ±5 trials Gaussian-smoothed running average. Because learning is defined as consistent above-average performance and not solely as overall high performance, we used a statistical expectation maximization algorithm to estimate when an ideal observer infers reliable learning from the succession of monkey choices in a block (Smith et al., 2004; see Methods). The ideal observer estimate showed that Monkey M reached learning, on average, after 7.9 trials (95% confidence range: 0.33) and Monkey S reached learning, on average, after 9.54 trials (95% confidence range: 0.45).

Overall performance level was stable across experimental sessions as the monkeys had learned the task structure during behavioral training sessions, which are not included in the analysis. Asymptotic performance measured across trials after initial learning was, on average, 87.3% correct (Monkey M: 89.1 ± 0.02%, Monkey S: 85.5 ± 0.02%; Figure 1C).

By task design, reversal blocks (i.e., block transitions where the color of the two stimuli was maintained but the color–reward association was reversed) represented 34% and 39% of the total number of blocks for Monkeys M and S, respectively. The average performance in these blocks was not significantly different from nonreversed blocks for Monkey M ($p > .05$, Mann–Whitney–Wilcoxon test) and only moderately different for Monkey S from 79.0% to 81.4% correct choices in reversed versus

nonreversed blocks ($p < .05$, Mann–Whitney–Wilcoxon test). In the following, we collapsed reversal and nonreversal block types as there were no major results qualitatively different across block types.

To validate quantitatively that the color dimension was the only feature used by animals to perform the task, we used a logistic regression analysis (see Methods). Sorting task features according to their subsequent predictive power for reward outcome through individual trials confirmed that stimulus colors were maximally explanatory of monkeys' behavior, whereas noncolor features had no systematic influence on the performance (Figure 2).

Evidence for an Optimal Internal Representation in the Learning of Feature Values

Having shown that animals were able to link choice outcomes (reward obtained from upward and downward saccades) to the feature that determined attentional selection, three questions arise that are addressed in the following: First, is there an optimal internal representation used to solve the task? Second, how are internal representations of feature values updated after experiencing outcomes? Third, is covert attentional selection fully described according to value-based mechanisms, or are there other non-value-based influences that systematically affect attentional performance?

To identify the computational processes that could control attentional selection, we devised and compared two Rescorla–Wagner type of RL models (Glimcher, 2011; Figure 3A and B). We describe one model, with a task set restricted to the relevant feature dimension (color) as feature-based RL, because it contains an internal model representation of only the relevant task features (Figure 3A). We contrasted this model with nonselective RL, which did not include any prior knowledge about which of the available decision variables were systematically linked to reward but rather relied on tracking values for all stimulus features that were available, including not only stimulus color but also location, rotation direction, and the time onset of rotations (Figure 3B).

Three independent criteria were combined to evaluate RL models (1) to account for the dynamics of learning in the task, (2) to analyze the plausibility of an RL mechanism for explaining monkey performance, and (3) to maximize the total number of trials in which monkey and model performance matched, corrected to penalize model biases against the least frequent outcome (see Methods). The direct comparison of RL models according to this evaluation revealed that the feature-based RL outperformed the nonselective RL in predicting covert attentional selection, evident in a significantly better (lower) optimized compound score of model performance in both monkeys (Figure 3C, feature-based vs. nonselective RL, comparison across 10 model realizations: Monkey M, $p < .005$; Monkey S, $p < .001$, Mann–Whitney–Wilcoxon test). The most prominent difference between models was that the stochastic selection process was considerably more deterministic (higher beta value) in the optimized nonselective RL model compared with the feature-based RL model (Figure 3D).

Despite the overall superiority of the feature-based RL model, the two models were indistinguishable in predicting the dynamics of learning within a block as inferred from the average monkey performance, and both models explained a similar proportion of animals' covert attentional selections in single trials (feature-based RL: Monkey M/S: 78.2%/72.2%, nonselective RL: Monkey M/S: 78.5%/71.5%; Figure 4A). The failure of nonselective RL became evident only when we compared the output of the stochastic selection process of the RL models with the selections made by the monkeys. Figure 4B illustrates that the probability to select the relevant stimulus according to the feature-based RL model closely followed the likelihood of correct choices made by the monkey (Figure 4B, top, the diagonal line represents a perfect match). In contrast, monkey choice likelihood deviated from the probability dictated by the nonselective RL model (Figure 4B, bottom). This result supports the suggestion that choices of the monkeys depend on prior covert attentional selection that operates on an internal representation of task-relevant feature space.

Erroneous Choices Reveal Non-value-based Selection Biases of Monkeys

The previous analysis showed that the probability of correct stimulus selections of the feature-based RL closely resembled the likelihood of correct overt choices of the monkeys on a trial-by-trial basis. This mechanism did not, however, explain why monkeys were committing nonrandomly distributed errors during asymptotic performance, that is, after they apparently had learned the reward predicting color.

To identify the source of these errors, we analyzed sequences of choices while at asymptotic performance and found that erroneous choices clustered together more often than expected by the performance of the feature-based RL model. Among all erroneous choices at peak behavior (M: 10.9%, S: 14.5%), consecutive errors made up 40.8% of errors for Monkey M and 34.3% for Monkey S (the proportional error patterns for monkey M [S]: 59.2% [65.7%] for CEC [correct–error–correct] successions, 24.2% [19.4%] for CEEC, 9.5% [9.0%] for CEEEC, etc.). Figure 5 illustrates how this error pattern deviated from a random distribution, revealing that both monkeys committed less errors in isolation and more errors in succession than expected for a stochastic error generating process. As might be expected given its stochastic selection mechanism, the feature-based RL system (and also the nonselective RL system) failed to capture this error pattern, both generating a pattern of errors close to random (Figure 5).

Value-based Attentional Selection Is Weighted by Non-value-based Selection Biases

The failure to account for the observed error pattern shows that feature-based RL must be complemented by additional mechanisms to explain the animal's attentional performance pattern. We thus extended the feature-based RL system and devised four additional models, each with a distinct mechanism for explaining behavior (see Methods). In particular, we tested the influence of (1) a direct effect of value-independent selection history onto feature-specific value representations (value history model, Figure 6A), (2) a hierarchical two-step selection process that incorporates an initial value-based feature selection as well as a subsequent value-independent input from selection history (hierarchical value-history model, Figure 6B), (3) a dynamic regulation of the selection stochasticity based on recent reward history (adaptive selection model, Figure 6C), and (4) in the last

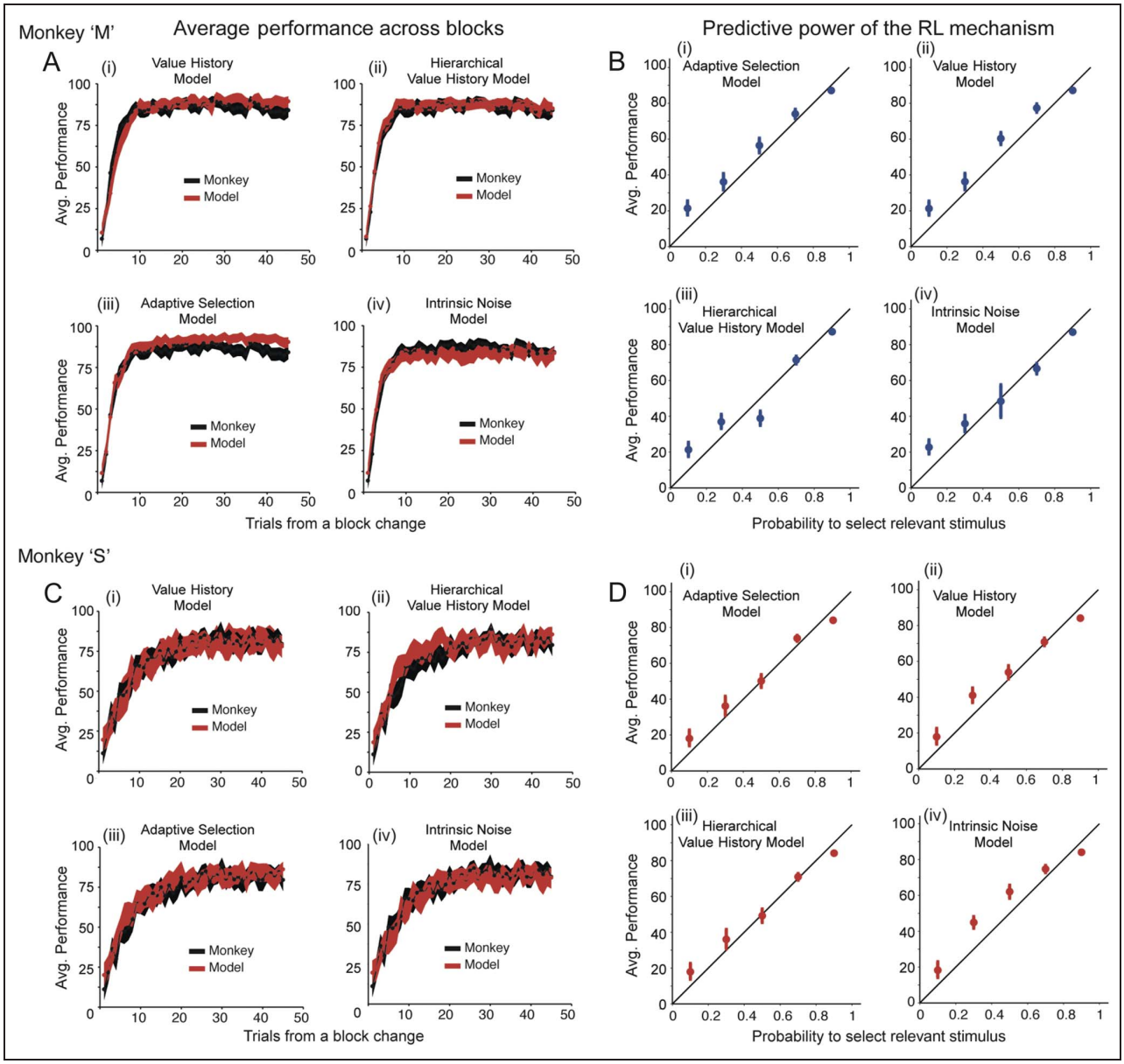


Figure 7. Average performance of monkeys and the four models that extend feature-based RL. (A and B) Results from Monkey M for (i) the value history model, (ii) the hierarchical value-history model, (iii) the adaptive selection model, and (iv) the intrinsic noise model, respectively. In A(i)–A(iv), the black shaded area shows the 95% CI around the mean for the monkey, in red for each model. B(i)–B(iv) show the averaged performance and 95% CI (error bars) of Monkey M (y axis) corresponding to five bins that fully span the range of the probability to select the relevant stimulus. A plausible model candidate requires the model's likelihood and monkey's performance to match each other. The degree to what this happens is quantified by the SSD. (C and D) Same as A and B for Monkey S.

Figure 8. The optimization scores and optimized parameters for extended models: the adaptive selection model, the hierarchical value-history model, the value history model, and the intrinsic noise model (see main text and Figure 6). Monkeys M and S results are shown in red and blue shaded cells, respectively.

	Model Types	Final Optimized Score	Optimization Measures			Optimized Parameters					
			SSD Mechanism	SSD Performance	Corr. % unexplained	α	β	β_H	γ	P_R	μ
M	Adaptive	0.116	0.011	0.016	0.320	0.3	-	3	-	-	0.06
	Hierarchical	0.134	0.001	0.065	0.341	0.5	1.6	-	0.5	-	-
	Value History	0.158	0.038	0.104	0.333	0.26	2.4	-	0.5	-	-
	Intrinsic Noise	0.236	0.005	0.337	0.367	0.36	1.7	-	-	0.27	-
S	Hierarchical	0.172	0.019	0.062	0.436	0.28	1.5	-	0.35	-	-
	Adaptive	0.177	0.03	0.076	0.426	0.23	-	2.4	-	-	0.12
	Value History	0.210	0.122	0.195	0.419	0.2	1.9	-	0.4	-	-
	Intrinsic Noise	0.25	0.249	0.089	0.42	0.16	1.95	-	-	0.3	-

model, we tested the influence of added noise to the system that is evenly distributed among choice features (intrinsic noise model, Figure 6D).

We optimized each of the four extended RL models separately using the compound criteria of model performance as before. Across models, the hierarchical value-history model and the adaptive selection model performed best (Figures 7 and 8). However, given that these models used additional parameters than the basic feature-based RL (Figure 3A), the improvement in explaining correct choices was at most marginal. However, in contrast to this marginal effect with respect to predicting correct choices, the prediction of erroneous choices separated model performance. In particular, predicting the pattern of consecutive errors revealed a clear advantage of the hierarchical value-history model against feature-based RL and each of the three remaining models in both monkeys (Figure 6E). Thus, the hierarchical value-history model closely predicted the error patterns evident in the two monkeys. It predicted the monkey's error patterns significantly better than the value-history model (for Monkey M [S]: $p < .001$ [.001], Kruskal–Wallis test), the adaptive selection model (for Monkey M [S]: $p < .001$ [.001], Kruskal–Wallis test), and the intrinsic noise model (for Monkey M [S]: $p < .001$ [.001], Kruskal–Wallis test). It is noteworthy that the prediction of error patterns was not an explicit criterion during optimization but emerged from the sequential (two-step) selection mechanism intrinsic in the hierarchical value-history model (Donoso et al., 2014; Ahn et al., 2008).

Analysis of Value-independent Selection Biases

Biases for stimulus features indicate limits for RL model predictions. Although our intrinsic noise model failed to improve the predictive abilities of the model-based RL framework, the distribution of errors observed in both monkeys at asymptotic behavior (Figures 5 and 6E) suggests that some non-value-based influence could affect behavior, either at the point of stimuli selection or somewhere else in the decision-making process (see below). To explore the role of possible biases relating to the selection of stimulus features that are independent of recent choice history, we ranked stimulus features according to the proportion in which they were associated to unrewarded trials (Figure 9). We found that both monkeys demonstrated an almost even distribution of errors across most of the task features indicating a similar likelihood to make a choice across task features independent of the task features' local associated value. Thus, errors because of non-value-based attentional biases did not represent a dominant behavioral strategy systematically used by the animals (Figure 2).

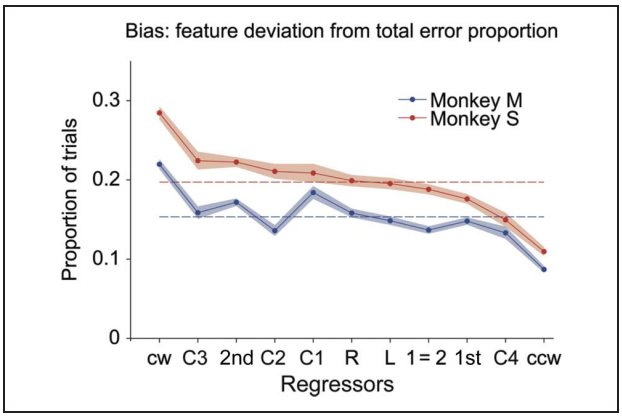


Figure 9. Non-value-based feature biases measured as the proportion of errors associated with each particular stimulus feature. Monkey S (red dashed line) presents an overall larger proportion of errors than Monkey M (red dashed line), and this pattern is systematic for each feature pair (solid lines). In principle, this could be because of non-value-based feature biases or explained by exploratory behavior. Given that colors are the features that predict attentional selection and eventually the behavioral choice (Figures 2 and 3), it might be expected that exploratory behavior would mainly stay within this dimension. However, we see that the proportion of errors is distributed among all features, with no clear preference for colors. Features are sorted from the highest to lowest feature biases according to Monkey S. Shaded areas represent 95% CIs.

DISCUSSION

We have shown that the learning of feature-based attentional selection in macaque monkeys can be predicted by models of RL with value-based selection mechanisms acting on a restricted feature space. Value-based learning explained the animals' behavior better when the updating of value representations was restricted to the feature dimension that was task relevant (color) and did not consider those feature dimensions (location, rotation direction, and time of rotation) that were not systematically linked to rewarding outcomes (in contrast to nonselective RL). This finding provides quantitative evidence in nonhuman primates that attentional selection can act on a task-specific representation of relevant features. Such feature representation can be formally described as internal state model within the RL framework. Implicit in this formulation is that attention is realized as a stochastic covert selection acting on feature-specific value predictions (Figure 10; Womelsdorf & Everling, 2015). A second main finding of our study is that the process of value-based attentional selection had to be complemented by an additional value-independent selection process to account for nonrandom influences of selection history on the pattern of erroneous choices.

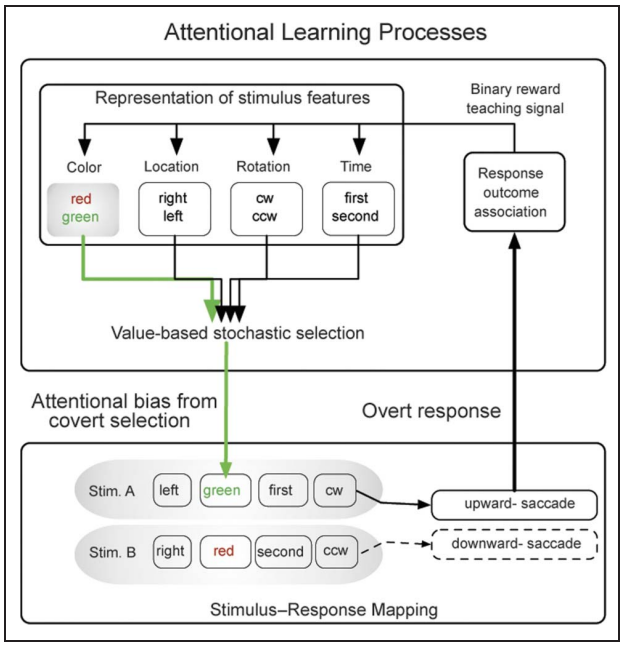


Figure 10. Separable processes underlying learning of attentional selection. Attentional selection relies on a covert decision-making process that is suitable to evaluate all stimulus features (i.e., color, location, etc.) but that, after practice, prioritizes the subset of relevant stimulus features that systematically link to reward (e.g., color dimension). In each trial, a particular covert attentional selection of a stimulus (green stimulus in the example) is established by value-based competition among elements in the task set representation. Values are then updated according to the response outcome. Note that the specific response outcome critically depends on a proper attentional selection to bias the relevant sensory processing (e.g., rotation discrimination) and, as a result, trigger an adequate sensory-response mapping (e.g., upward saccadic response to report a clockwise rotation).

Internal Value Predictions of Task-relevant Features Provide the Reference for Attention

We found that value-based learning of attention is not naive with respect to features in the environment that are systematically linked to reward. When an animal received a rewarding outcome, this reward was linked in individual trials to a choice on a particular rotation direction (clockwise or counterclockwise), a particular time onset (e.g., first or second), a particular location (left or right), and color (e.g., blue) of the stimulus. If all these task variables were considered equivalently while updating value representations, the nonselective RL controller would have outperformed the feature-based learner, as local correlations of nonrelevant features with reward outcomes would have impacted on monkeys' behavior. Instead, these multiple features were not treated equally in the credit assignment process (Figure 4B). The updating of values was better described as being selective to prioritized task-specific representations. This finding highlights the idea that a key component of flexible attention lies in the evaluation process of how causal sources of outcomes are identified and credited for producing the outcome. This empirically derived conclusion supports previous modeling studies that implicate attentional selection signals as critical gating signals for plasticity and learning of task-relevant sensory features (Rombouts, Bohte, Martinez-Trujillo, & Roelfsema, 2015; Alexander, 2007; Roelfsema & van Ooyen, 2005; see also Roelfsema, van Ooyen, & Watanabe, 2010). In summary, our findings show that the deployment of attention can be efficiently adjusted according to feature–reward associations. We should note that we could not model the origins of the segmentation of task-relevant variables in the current data set that was limited to later stages of task learning. However, we believe that it is an important future task to extend the feature-based RL model to include the learning of a segmentation between task-relevant and task-irrelevant features by processes using either meta-learning mechanisms (Ardid, Balcarras, & Womelsdorf, 2014; Gershman & Niv, 2010) or, for example, by adding an independent slow learning process that tracks input statistics and derives policies from it (Legenstein, Wilbert, & Wiskott, 2010).

Attentional Flexibility versus Stickiness

After steep reversal learning, the performance of monkeys did not reach optimality, but rather, animals continued to make wrong, unrewarded choices in 10–15% of all trials during a

period where expected values for stimulus color were at a constant high level. We found that this 10–15% failure rate can be traced back to three identifiable sources that are informative about the processes controlling attention. The largest proportion of errors was accounted for by the softmax stochastic selection process (through the β parameter) that imposed a nonzero probability to select the stimulus features with the lowest values. This aspect is important because it supports the notion that attention can be conceptualized as a stochastic selection process similar to conceptualization of overt (motor) choice (Gottlieb, 2012; Rangel & Hare, 2010).

A second source of errors in our task are feature biases of the animals that are independent of fluctuations in value predictions and reflect "default" tendencies of animals choices (see Figure 9), although the animals could not (and did not) systematically deploy such simple strategies to solve our task (Shteingart & Loewenstein, 2014; Figure 2).

The third source of erroneous performance referred explicitly to the pattern of errors that deviated from a purely stochastic process once in an asymptotic regime, with an evident tendency to repeat erroneous (unrewarded) choices (Figures 5 and 6). Both animals showed this deviation from random error generation, resembling perseveration tendencies and habit intrusions known from the motor domain. However, the repeated errors in our task referred to repetitions of the attentional selection (i.e., based on color) from the previous trial. Only a single model was able to capture this error pattern by means of a sequential (two-step) process that complemented the basic value-based selection with a second selection process that pushed the final overt choice toward the previous selection.

Such a weighting of a current trial's value-based selection is in fact an efficient strategy when the previous selection was rewarded; hence, repeating the same attentional selection is, in such a condition, a strategy that reduces effort and costs (Shenhav, Botvinick, & Cohen, 2013). However, when the previous trial's covert choice was an error and led to no reward, weighting the current value-based selection toward the nonrewarded previous covert choice is detrimental and incurs costs. This cost of committing two or more consecutive errors represented a substantial subproportion of error trials (34–41% of the 10–15% total number of errors in the task), which may relate to the actual cost the animals are able to tolerate in the control of attention, given the effort it would take for them to improve performance. This interpretation is consistent with a recent proposal that quantifies the expected value of (attentional) control by estimating the (sum of anticipated) payoffs against the costs to establish sufficiently strong control to obtain such payoffs (Shenhav et al., 2013). According to this interpretation, the cost of committing errors in our task is traded against the level of effort (i.e., strength of control) that would be required to improve performance (number of rewarded trials). In particular, in our task, improving performance requires constant updating of feature value representations and covert stimulus selection. We can thus speculate that the hypothesized quantity about the expected control intensity is related to the γ parameter in our hierarchical value-history model, which is adjusted to each monkey's trade-off between effort and payoff. The lower this parameter, the higher is the effort to receive more value-based payoffs, and on the contrary, the higher the parameter, the larger is the attentional stickiness and the tendency to perseverate on previous attentional selections.

Implications for Models of Attention: Value- and Non-value-based Processes

The success in explaining actual attentional learning in primates with a feature-based RL mechanism that is weighted with an attentional stickiness process has further implications for theories of attention.

First, the results suggest that the valuation system plays a key role to determine to what features selective attention is shifted to, independently of the saliency of those features (Womelsdorf & Everling, 2015; Krauzlis, Bollimunta, Arcizet, & Wang, 2014; Chelazzi et al., 2013; Kaping et al., 2011; Tatler et al., 2011; Navalpakkam, Koch, Rangel, & Perona, 2010). Value representations in the RL framework are predictions of stimulus values (predictions of outcomes), demonstrating that the covert control of visual attention can be understood from a predictive coding perspective such that feature value predictions resemble reward value predictions in the domain of overt goal-directed behavior, decision-making, and planning (van der Meer, Kurth-Nelson, & Redish, 2012; Wilson & Niv, 2011; Seymour & McClure, 2008; Dehaene & Changeux, 2000). This conclusion resonates well with studies documenting the influence of expectations for visual perception and perceptual inferences (Series & Seitz, 2013; Summerfield & Egner, 2009), the influence of secondary reward associations to modify basic visual search efficiency (Anderson et al., 2011), and a growing literature documenting the influence of actual attentional experiences to shape reward memories and attentional priorities through learning mechanisms (Gottlieb, Hayhoe, Hikosaka, & Rangel, 2014; Chelazzi et al., 2013; Awh, Belopolsky, & Theeuwes, 2012; Della Libera & Chelazzi, 2009).

Second, attention in our task also depends on a second process that weights the value-based selection based on repeating the selection of previous trials irrespective of whether that selection was rewarded or unrewarded. Such a reward-insensitive mechanism is particularly useful in probabilistic choice contexts where the lack of reward at one occasion can be a mere stochastic event that is better ignored to maximize reward intake in the long run (Lau & Glimcher, 2005). In our task with a deterministic reward schedule within blocks of trials, the weighting of the current choice toward previous choices

is reminiscent of (1) previous trial effects in stimulus–response learning tasks (Fecteau & Munoz, 2003) and shares similarity with (2) habitual stimulus response control (Dolan & Dayan, 2013), (3) habit intrusions (de Wit et al., 2012), (4) behavioral perseverations and stickiness (Dayan & Berridge, 2014; Huys et al., 2011), and (5) intertrial priming and repetition memory effects (Anderson, 2013; Awh et al., 2012; Kristjansson & Campana, 2010; Kristjansson, 2006). All these listed effects are empirical demonstrations of the apparent influence of a memory of recent choices and attentional selections on current attentional performance. Whether these various history and memory effects serve as primary controllers of attentional selections or should better be conceived of as modulators of attention will be a question for future research. Our findings are more supportive of the former suggestion, revealing that selection history influences attentional performance in such a dominant way that it should be considered a separate control process underlying attentional selection, which complements value-based control.

Conclusion and Outlook

Taken together, we have illustrated a formal framework of attentional selection in nonhuman primates that provides explicit and testable hypotheses about the specific subprocesses underlying attentional control. Our hierarchical value-history RL model specifies these three main attentional subprocesses as (1) the feature-specific learning of value predictions, (2) the stochastic value-based selection process, and (3) a non-value-based memory bias that drives the system toward previously selected information. We speculate that the very structures implicated in stimulus valuation, RL, and decision-making are key structures in controlling the focus of visual attention (Womelsdorf & Everling, 2015). Each of these processes is possibly associated with separable neuronal circuits in the primate prefrontal, striatal, and medial-temporal lobe systems. Circuits within prefrontal regions presumably include the lateral pFC, an area that may not have an anatomical and functional analog in the nonprimate brain (Passingham & Wise, 2012). Our study in nonhuman primates could thus become a versatile starting point to understand how multiple choice systems and subprocesses underlying stimulus selection interact to control attention in primates.

Acknowledgments

This work was supported by grants from the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Ontario Ministry of Economic Development and Innovation. We thank Johanna Stucke for her help with assisting with animal training and care. We thank anonymous reviewers for helpful comments on earlier versions of this manuscript.

Reprint requests should be sent to Dr. Thilo Womelsdorf, Dr. Salva Ardid, or Dr. Matthew Balcarras, Department of Biology, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada, or via e-mail: thiwom@yorku.ca, sardid@bu.edu, mbalcarr@yorku.ca.

REFERENCES

Ahn, W. Y., Busemeyer, J. R., Wagenmakers, E. J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science, 32,* 1376–1402.

Alexander, W. H. (2007). Shifting attention using a temporal difference prediction error and high-dimensional input. *Adaptive Behavior, 15,* 121–133.

Anderson, B. A. (2013). A value-driven mechanism of attentional selection. *Journal of Vision, 13,* 1–16.

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences, U.S.A., 108,* 10367–10371.

Ardid, S., Balcarras, M., & Womelsdorf, T. (2014). "Adaptive learning" as a mechanistic candidate for reaching optimal task-set representations flexibly. *BMC Neuroscience, 15,* P8.

Ardid, S., & Wang, X. J. (2013). A tweaking principle for executive control: Neuronal circuit mechanism for rule-based task switching and conflict resolution. *Journal of Neuroscience, 33,* 19504–19517.

Asaad, W. F., & Eskandar, E. N. (2008). A flexible software tool for temporally-precise behavioral control in Matlab. *Journal of Neuroscience Methods, 174,* 245–258.

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top–down versus bottom–up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences, 16,* 437–443.

Cai, X., & Padoa-Schioppa, C. (2014). Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. *Neuron, 81,* 1140–1151.

Chelazzi, L., Perlato, A., Santandrea, E., & Della Libera, C. (2013). Rewards teach visual selective attention. *Vision Research, 85,* 58–72.

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

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective & Behavioral Neuroscience, 14,* 473–492.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience, 3(Suppl.),* 1218–1223.

de Wit, S., Watson, P., Harsay, H. A., Cohen, M. X., van de Vijver, I., & Ridderinkhof, K. R. (2012). Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. *Journal of Neuroscience, 32,* 12066–12075.

Dehaene, S., & Changeux, J. P. (2000). Reward-dependent learning in neuronal networks for planning and decision making. *Progress in Brain Research, 126,* 217–229.

Della Libera, C., & Chelazzi, L. (2009). Learning to attend and to ignore is a matter of gains and losses. *Psychological Science, 20,* 778–784.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society, Series B: Methodological, 39,* 1–38.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80,* 312–325.

Donoso, M., Collins, A. G., & Koechlin, E. (2014). Human cognition. Foundations of human reasoning in the prefrontal cortex. *Science, 344,* 1481–1486.

Fecteau, J. H., & Munoz, D. P. (2003). Exploring the consequences of the previous trial. *Nature Reviews Neuroscience, 4,* 435–443.
Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology, 20,* 251–256.
Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences, U.S.A., 108(Suppl. 3),* 15647–15654.
Gottlieb, J. (2012). Attention, learning, and the value of information. *Neuron, 76,* 281–295.
Gottlieb, J., Hayhoe, M., Hikosaka, O., & Rangel, A. (2014). Attention, reward, and information seeking. *Journal of Neuroscience, 34,* 15497–15504.
Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P., & Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences, U.S.A., 108,* 18120–18125.
Huys, Q. J., Cools, R., Golzer, M., Friedel, E., Heinz, A., Dolan, R. J., et al. (2011). Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLoS Computational Biology, 7,* e1002028.
Kaping, D., Vinck, M., Hutchison, R. M., Everling, S., & Womelsdorf, T. (2011). Specific contributions of ventromedial, anterior cingulate, and lateral prefrontal cortex for attentional selection and stimulus valuation. *PLoS Biology, 9,* e1001224.
Kennerley, S. W., Behrens, T. E., & Wallis, J. D. (2011). Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nature Neuroscience, 14,* 1581–1589.
Krauzlis, R. J., Bollimunta, A., Arcizet, F., & Wang, L. (2014). Attention as an effect not a cause. *Trends in Cognitive Sciences, 18,* 457–464.
Kristjansson, A. (2006). Rapid learning in attention shifts: A review. *Visual Cognition, 13,* 324–362.
Kristjansson, A., & Campana, G. (2010). Where perception meets memory: A review of repetition priming in visual search tasks. *Attention, Perception, & Psychophysics, 72,* 5–18.
Kruschke, J. K., & Hullinger, R. A. (2010). Evolution of attention in learning. In N. A. Schmajuk (Ed.), *Computational models of conditioning* (pp. 10–52). Cambridge: Cambridge University Press.
Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior, 84,* 555–579.
Lau, B., & Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron, 58,* 451–463.
Legenstein, R., Wilbert, N., & Wiskott, L. (2010). Reinforcement learning on slow features of high-dimensional input streams. *PLoS Computational Biology, 6.*
Luk, C. H., & Wallis, J. D. (2013). Choice coding in frontal cortex during stimulus-guided or action-guided decision-making. *Journal of Neuroscience, 33,* 1864–1871.
Navalpakkam, V., Koch, C., Rangel, A., & Perona, P. (2010). Optimal reward harvesting in complex perceptual environments. *Proceedings of the National Academy of Sciences, U.S.A., 107,* 5232–5237.
Padoa-Schioppa, C. (2011). Neurobiology of economic choice: A good-based model. *Annual Review of Neuroscience, 34,* 333–359.
Passingham, R. E., & Wise, S. P. (2012). *The neurobiology of the prefrontal cortex: Anatomy, evolution, and the origin of insight*. Oxford: Oxford University Press.
Peck, C. J., Jangraw, D. C., Suzuki, M., Efem, R., & Gottlieb, J. (2009). Reward modulates attention independently of action value in posterior parietal cortex. *Journal of Neuroscience, 29,* 11182–11191.
Peck, C. J., Lau, B., & Salzman, C. D. (2013). The primate amygdala combines information about space and value. *Nature Neuroscience, 16,* 340–348.
Rangel, A., & Clithero, J. A. (2014). The computation of stimulus values in simple choice. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics: Decision making and the brain* (2nd ed., pp. 125–148). Amsterdam/Boston: Elsevier/Academic Press, Academic Press is an imprint of Elsevier.
Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology, 20,* 262–270.
Roelfsema, P. R., & van Ooyen, A. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural Computation, 17,* 2176–2214.
Roelfsema, P. R., van Ooyen, A., & Watanabe, T. (2010). Perceptual learning rules based on reinforcers and attention. *Trends in Cognitive Sciences, 14,* 64–71.
Rombouts, J. O., Bohte, S. M., Martinez-Trujillo, J., & Roelfsema, P. R. (2015). A learning rule that explains how rewards teach attention. *Visual Cognition, 23,* 179–205.
Rushworth, M. F., & Behrens, T. E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience, 11,* 389–397.
Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron, 70,* 1054–1069.
Series, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience, 7,* 668.
Seymour, B., & McClure, S. M. (2008). Anchors, scales and the relative coding of value in the brain. *Current Opinion in Neurobiology, 18,* 173–178.
Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron, 79,* 217–240.
Shteingart, H., & Loewenstein, Y. (2014). Reinforcement learning and human behavior. *Current Opinion in Neurobiology, 25C,* 93–98.
Smith, A. C., & Brown, E. N. (2003). Estimating a state-space model from point process observations. *Neural Computation, 15,* 965–991.
Smith, A. C., Frank, L. M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., et al. (2004). Dynamic analysis of learning in behavioral experiments. *Journal of Neuroscience, 24,* 447–461.
Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science, 304,* 1782–1787.
Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences, 13,* 403–409.
Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11,* 5.
Tsotsos, J. K. (2011). *A computational perspective on visual attention* (1st ed.). Cambridge, MA: MIT Press.
van der Meer, M., Kurth-Nelson, Z., & Redish, A. D. (2012). Information processing in decision-making systems. *The Neuroscientist, 18,* 342–359.
Wilson, R. C., & Niv, Y. (2011). Inferring relevance in a changing world. *Frontiers in Human Neuroscience, 5,* 189.
Wirth, S., Yanike, M., Frank, L. M., Smith, A. C., Brown, E. N., & Suzuki, W. A. (2003). Single neurons in the monkey hippocampus and learning of new associations. *Science, 300,* 1578–1581.
Womelsdorf, T., & Everling, S. (2015). Long-range attention networks: Circuit motifs underlying endogenously controlled stimulus selection. *Trends in Neurosciences*. doi:10.1016/j.tins.2015.08.009.
Wunderlich, K., Rangel, A., & O'Doherty, J. P. (2010). Economic choices can be made using only stimulus values. *Proceedings of the National Academy of Sciences, U.S.A., 107,* 15005–15010.