Your favorite color makes learning more adaptable and precise

Shiva Farashahi[1]*, Katherine Rowe[1]*, Zohra Aslami[1], Daeyeol Lee[2-4], Alireza Soltani[1]

[1]*Department of Psychological and Brain Sciences, Dartmouth College, NH, 03784*

[2]*Department of Neuroscience, Yale University School of Medicine, New Haven, CT, 06520*

[3]*Kavli Institute for Neuroscience, Yale University School of Medicine, New Haven, CT, 06520*

[4]*Department of Psychology, Yale University, New Haven, CT, 06520*

** These authors contributed equally to this work*

Correspondence: A.S., Department of Psychological and Brain Sciences, HB 6207, Dartmouth College, Hanover, NH, 03755. Email: soltani@dartmouth.edu

Manuscript information: 51 pages; 8 figures; and 4 tables.

Abstract

Learning from reward feedback is essential for survival but can become extremely challenging when choice options have multiple features and feature values (curse of dimensionality). Here, we propose a general framework for learning reward values in dynamic multi-dimensional environments via encoding and updating the average value of individual features. We predicted that this feature-based learning occurs not just because it can reduce dimensionality, but more importantly because it can increase adaptability without compromising precision. We experimentally tested this novel prediction and found that in dynamic environments, human subjects adopted feature-based learning even when this approach does not reduce dimensionality. Even in static low-dimensional environment, subjects initially tended to adopt feature-based learning and switched to learning individual option values only when feature values could not accurately predict all objects values. Moreover, behaviors of two alternative network models demonstrated that hierarchical decision-making and learning could account for our experimental results and thus provides a plausible mechanism for model adoption during learning in dynamic environments. Our results constrain neural mechanisms underlying learning in dynamic multi-dimensional environments, and highlight the importance of neurons encoding the value of individual features in this learning.

Introduction

Human behavior is marked by a sophisticated ability to attribute reward outcomes to appropriate choices and events with surprising nuance. Learning from reward feedback is essential for survival, and understanding the mechanisms underlying this ability is a major focus of cognitive neuroscience. In nature and in ecologically valid laboratory settings, however, such learning is challenging because choices have many features (e.g. color, shape, texture), each of which can take different values, resulting in a large number of options for which reward values have to be learned. This is referred to as the "curse of dimensionality," because standard models of learning do not scale up with the increase in dimensionality and thus the number of possible options in the environment (Barto & Mahadevan, 2003; Diuk, Tsai, Wallis, Botvinick, & Niv, 2013; Hastie, Tibshirani, & Friedman, 2001; Sutton & Barto, 1998).

An increase in dimensionality creates two main difficulties for the standard reinforcement learning (RL) models that try to directly learn the value of individual options (i.e. model-free approach). First, learning is too slow and imprecise due to the many reward outcomes needed to get an accurate estimate of the reward value, as well as the possibility that reward contingencies might quickly change over time. Second, the value of certain options not yet encountered cannot be estimated (Kahnt, Chang, Park, Heinzle, & Haynes, 2012). Thus, two approaches are commonly proposed to overcome the curse of dimensionality. One approach is to construct a simplified representation of the stimuli and therefore, to learn only a small subset of features and ignore others (Niv et al., 2015; Wilson & Niv, 2012). However, there are behavioral and neural data suggesting that human subjects process all features of each option, rather than focusing on a single informative dimension (Wunderlich, Beierholm, Bossaerts, & O'Doherty, 2011). Moreover, ignoring subsets of features could be detrimental in dynamic environments where previously non-informative features can suddenly become informative. The other approach is to infer the structure of the task and create rules to estimate reward values of options based on their features, a process often referred to as the model-based approach, which requires a much smaller set of values to be learned (Braun, Mehring, & Wolpert, 2010; Dayan & Berridge, 2014; Gershman & Niv, 2010; Maia, 2009).

A simple form of the model-based approach is feature-based learning, in which the average values of all features are learned and then combined to estimate the reward values for individual options. In addition to reducing dimensionality without ignoring any features, feature-based learning also enables faster learning and more adaptability without compromising precision. This is possible because reward values of all features of the selected option can be updated based on a single reward feedback. In contrast, simply increasing the learning rates in a model-free approach can improve adaptability but also adds noise in the estimation of reward values for individual options (adaptability-precision tradeoff). Therefore, by enabling faster learning without sacrificing precision, feature-based learning can ameliorate the adaptability-precision tradeoff. Feature-based learning can be implemented neuronally using a relatively small number of value-encoding neurons with pure feature selectivity, i.e. neurons that represent the reward value in a single dimension, such as color or shape. In contrast, object-based learning in a model-free approach requires myriad mixed selectivity neurons tuned to specific combinations of various features. This difference makes feature-based learning computationally more feasible than object-based learning. However, feature-based learning is only beneficial if a generalizable set of rules exists such that the reward value of all options can be accurately constructed from their feature values.

Therefore, the main advantage of feature-based learning might be to overcome the adaptability-precision tradeoff. To test this hypothesis, we constructed a general framework and designed a series of experiments to explore and measure how multiple factors encourage the adoption of feature-based versus object-based learning. Moreover, we designed and tested two alternative network models in order to capture our experimental observations.

Materials and Methods

General framework. We constructed a general framework for model adoption during learning of reward values in a dynamic, multi-dimensional environment. Assuming that objects (options) contain m features, each of which can take one of n feature values, there are n^m objects/options in the environment. In our simulations, the probability of reward on individual objects (the reward

matrix) was constructed by first assigning a set of equally-spaced likelihood ratios (*LR*) to *n* possible values of each feature in all *m* dimensions. The minimum and maximum values of *LR*'s were set to *1/x* and to *x* ($x > 1$), respectively. For example, for $n = 3$, $LR(F_{ij}) = \{1/2, 1, 2\}$ where F_{ij} is the feature value *j* ($j = 1,\ldots,n$) of feature *i* ($i = 1,\ldots,m$). The *LR* for a given object *a* ($LR(O_a)$) was assigned by multiplying the *LR*'s of all features of that object: $LR(O_a) = \prod_{i=1, for\ F_{ij}\ present\ in\ O_a}^{m} LR(F_{ij})$. The probability of reward on each object was then computed by transforming the object's *LR* to the probability of reward: $p_r(O_a) = LR\ (O_a)/(1 + LR\ (O_a))$.

To calculate the estimated reward matrix based on the average feature values, we first computed the average reward value for each value of a feature (e.g. red, green, triangles, squares) by averaging the reward values of objects that contain that feature: $p_r(F_i) = (1/n^{m-1}) \sum_{O_a\ contains\ F_{ik}} p_r(O_a)$. By combining these average feature values we then generated an estimated reward value, $\tilde{p}_r(O_a)$, using the Bayes theorem:

$$\tilde{p}_r(O_a) = (p_r(F_{1j}) \times p_r(F_{2k}) \times \ldots)/(p_r(F_{1j}) \times p_r(F_{2k}) \times \ldots + (1 - p_r(F_{1j})) \times (1 - p_r(F_{2k})) \times \ldots) \quad \text{for } O_a \text{ containing of } F_{1j}, F_{2k}, \text{ etc.} \qquad \text{(Eq. 1)}$$

By randomly shuffling elements of the reward matrix in all dimensions except one (which we call the informative feature), we generated environments with varying values of correlation between the reward matrix and estimated reward matrix (i.e. correlation between the reward value of options and the estimated reward value of options based on their average feature values). This correlation was used as the generalizability index, which can take on any value between -1 and 1. Without any shuffling, we get a fully generalizable environment (generalizability index equal to 1) where the reward values of all options can be precisely constructed from the average reward values of their features.

The task for the decision maker is to learn the reward value of options via reward feedback in order to choose between two alternative options on each trial. The model with object-based learning directly learns the reward value of all objects based on reward feedback on each trial:

$$V_{O_a}(t+1) = V_{O_a}(t) + \alpha\left(1 - V_{O_a}(t)\right), \quad if\ R(t) = 1$$

$$V_{O_a}(t+1) = V_{O_a}(t) - \alpha\left(V_{O_a}(t)\right), \quad if\ R(t) = 0 \qquad \text{(Eq. 2)}$$

where t represents the trial number, $V_{O_a}(t)$ is the reward value of the chosen object, $R(t)$ is the trial outcome (1 for rewarded, 0 for unrewarded), and α is the learning rate. The value of the unchosen object is not updated. The model with feature-based learning learns the average reward value of individual feature values (e.g. red, green, triangles, squares), $V_{F_i}(t)$, using the same update rule as in Equation 2, but applying to all features of the chosen object. Therefore, the object-based learning model only updates one value function after each feedback whereas the feature-based learning model updates the average reward value of all features of the chosen object.

To measure how well a model based on object-based or feature-based learning can differentiate between different options at a given point in time, we defined the differential signal, $S_O(t)$, in the object-based learning model as follows:

$$S_O(t) = \frac{1}{n^m \times (n^m - 1)} \sum_{i=1}^{n^m} \sum_{j=1}^{n^m} \left(V_{O_i}(t) - V_{O_j}(t)\right) sign(p_r(O_i) - p_r(O_j)) \qquad \text{(Eq. 3)}$$

where $p_r(O_i)$ is the probability of reward on object i. The differential signal for the feature-based learning model, $S_F(t)$, was computed by replacing $V_{O_i}(t)$ in the above equation with the estimated reward value $\tilde{V}_{O_i}(t)$, which was computed by replacing $p_r(F_i)$ in Equation 1 with $V_{F_i}(t)$. Therefore, the differential signal measures how reward values estimated by a given model correctly differentiate between object values. By comparing the time course of differential signal for the object-based and feature-based learning models (using the same learning rate), we computed the time at which the object-based learning model carries a stronger differential signal than the feature-based learning model (the 'switch point'). A larger switch point indicates the superiority (better performance) of the feature-based relative to the object-based learning model for a longer amount of time, whereas zero switch point indicates that the object-based learning model is always better.

Subjects. Subjects were recruited from the Dartmouth College student population. In total, 59 subjects were recruited (34 females) to perform the choice task in Experiment 1 or 2 or both (33 subjects performed in both experiments). These resulted in N = 51 and N = 41 sets of behavioral

data for Experiments 1 and 2, respectively. To exclude subjects whose performances were not significantly different from chance (0.5), we used a performance threshold of 0.5268 (equal to 0.5 plus 3 times s.e.m., based on the average of 768 trials in each session of Experiments 1 and 2). This resulted in the exclusion of data from 5 of 51 sets in Experiment 1, and 14 of 41 sets in Experiment 2. The data from the remaining 73 sessions was used for further analysis (N = 46 and N = 27 for Experiments 1 and 2, respectively). For Experiment 3, 36 additional subjects were recruited (20 females) and a performance threshold of 0.5317 (equal to 0.5 plus 3 times s.e.m., based on the average of 560 trials) was used to exclude subjects whose performance was indistinguishable from chance (N = 4). In total, only two subjects participated in all three experiments, and this occurred over four months. For Experiment 4, 36 new subjects were recruited (22 females) and a performance threshold of 0.5289 (equal to 0.5 plus 3 times s.e.m., based on the average of 672 trials) was used to exclude subjects whose performance was indistinguishable from chance (N = 9). No subject had a history of neurological or psychiatric illness. Subjects were compensated with a combination of money and "t-points," which are extra-credit points for classes within the Dartmouth College Psychological and Brain Sciences department. The base rate for compensation was $10/hour or 1 t-point/hour. Subjects were then additionally rewarded based on their performance, by up to $10/hour. All experimental procedures were approved by the Dartmouth College Institutional Review Board, and informed consent was obtained from all subjects before participating in the experiment.

Experiments 1 and 2. In these experiments, subjects completed two sessions (each session composed of 768 trials and lasting about one hours) of a choice task during which they selected between a pair of objects on each trial (Fig.1A). Objects were one of four colored shapes: blue triangle, red triangle, blue square, and red square. Subjects were asked to choose the object that was more likely to provide a reward in order to maximize the total number of reward points, which would be converted to monetary reward and/or t-points at the end of the experiment.

[Figure 1 about here]

In each trial, the selection of an object was rewarded only according to its reward probability and independently of the reward probability of the other object. This reward schedule was fixed for a block of trials (block length, $L = 48$), after which it changed to another reward schedule without

any signal to the subject. Sixteen different reward schedules were used (all permutations of four reward probabilities [0.1, 0.3, 0.7. 0.9]), eight of which consisted of generalizable rules for how combinations of feature values (color or shape) predicted reward probability for different objects (Fig.1B). This set of reward schedules was intended to allow for generalization across features. The other eight schedules lacked generalizable rules for how combinations of feature values predicted reward probability for different objects (Fig.1C). For example, the schedule notated as 'Rs' corresponds to red objects being much more rewarding than blue objects, square objects being more rewarding than triangle objects, and color ('R' comes first) being more informative than shape ('s' comes second). In this generalizable schedule, red square was the most rewarding object whereas blue triangle was the least rewarding object. For non-generalizable schedules, only one of the two features was on average informative of reward values. For example, the 'r1' schedule indicated that, overall, red objects were slightly more rewarding than blue objects, but there was no generalizable relationship between the rewarding values of individual objects (e.g. red square was the most rewarding object, but red triangle was less rewarding than blue triangle). In other words, the non-generalizable set of reward schedules was designed so that a rule based on feature combination could not predict reward probability on all objects. For example, learning something about a red triangle did not necessarily tell the subject anything about other red objects or other triangle objects.

The main difference between Experiments 1 and 2 was that the environments in these experiments were composed of reward schedules with a generalizable and non-generalizable rule, respectively. Critically, as the subjects moved between blocks of trials, reward probabilities for the informative features (e.g., red and blue for the switch between Rs and Bs) were reversed without any changes in the average reward probabilities for the non-informative features. This was done without any cue to the subject in order to induce volatility in reward information throughout the task (Fig.1E, G). In addition, the average reward probabilities for the non-informative feature changed (e.g., from Bs and Rs to Bt and Rt) every four blocks (super-blocks; Fig.1D, E). Generalizable and non-generalizable reward schedules were used to create two separate environments. On each block of the generalizable environment (Experiment 1), feature-based rules could be used to correctly predict reward throughout the task (Fig.1B). In the non-generalizable environment (Experiment 2), however, the reward probability assigned to each object could not be determined based on the objects' features (e.g. in an additive fashion) and

instead depended on the combination of features (see above; Fig.1C). Each subject performed the experiment in each environment once, and either color or shape was more informative for both environments. The more informative feature was randomly assigned and counter-balanced across subjects to minimize the effects of intrinsic color or shape biases.

Experiment 3. In this experiment, subjects completed two sessions, each of which included 280 choice trials interleaved with five or eight short blocks of estimation trials (each block with eight trials). On each trial of the choice task, the subject was presented with a pair of objects and was asked to choose the object that they believed would provide the most reward. These objects were drawn from a set of eight objects, which were constructed using combinations of three distinct patterns and three distinct shapes (Fig.2A-B; one of nine possible objects with reward probability of 0.5 was excluded to shorten the duration of the experiment). The three patterns and shapes were selected randomly for each subject. The two objects presented on each trial always differed in both pattern and shape. Other aspects of the choice task were similar to those in Experiments 1 and 2, except that reward feedback was given for both objects rather than just the chosen object, in order to accelerate the learning. During the estimation blocks, subjects provided their estimates of the probability of reward for individual objects. Possible values for these estimates were from 5% to 95%, in 10% increments (Fig.2E). All subjects completed five blocks of estimation trials throughout the task (after trials 42, 84, 140, 210, and 280 of the choice task), but some subjects had three additional blocks of estimation trials (after trials 21, 63, and 252) to better assess the estimations over time. Each session of the experiment was about 45 minutes in length, with a break before the beginning of the second session. The second session was similar to the first, but with different sets of shapes and patterns.

[Figure 2 about here]

Selection of a given object was rewarded (independently of the other object on a given trial) based on a reward schedule with a moderate level of generalizability such that reward probability of some individual objects could not be determined based on their feature values. Because of the larger number of objects, the reward schedule was more complex than that used in Experiment 1 but did not change over the course of the experiment. Non-generalizable reward matrices can be constructed in many ways. In experiment 3, one feature (shape or pattern) was informative about

reward probability while the other feature was not. Although the informative feature (e.g. pattern in Fig.2A and shape in Fig.2B) was on average predictive of reward, this prediction was not generalizable. That is, there were objects that contained the most rewarding feature value (e.g. S1P3 in Fig.2A) but were less rewarding than objects that did not contain this feature (e.g. S1P2). Finally, the non-informative feature did not follow any generalizable rule such that the average reward probability across objects with similar feature value (e.g. S1P1, S1P2, S1P3 in Fig.2A) was 0.5. This reward schedule ensured that subjects would not be able to use a generalizable feature-based rule to accurately predict reward probability for all objects. Similarly to Experiments 1 and 2, the informative feature was randomly assigned and counter-balanced across subjects to minimize the effects of intrinsic pattern or shape biases.

Experiment 4. This experiment was similar to Experiment 3, except that we used four feature values for each feature (shape and pattern) resulting in an environment with a higher dimensionality. Each subject completed two sessions, each of which included 336 choice trials interleaved with five or eight short blocks of estimation trials (each block with eight trials). The objects in this experiment were drawn from a set of twelve objects, which were combinations of four distinct patterns and four distinct shapes (Fig.2C-D; four of sixteen possible objects with reward probability 0.5 were removed to shorten the duration of the experiment). The four patterns and shapes were selected randomly for each subject. The probabilities of reward on different objects (reward matrix) were set such that there was one informative feature, and the minimum and maximum average reward value for features were similar for Experiments 3 and 4.

Data analysis. We utilized two methods to test how subjects determined the values of objects using their reward probability estimates. First, we used linear regression to fit the estimates of reward probabilities as a function of the actual reward probabilities assigned to each object (object-based term), the reward probabilities estimated based on average feature values using the Bayes theorem (Eq.1) (feature-based term), and a constant. The constant (bias) in this regression model quantifies subjects' overall bias in reporting reward values. Moreover, the relative weight of bias to the regression coefficient for the feature-based term indicates the subject's lack of discrimination between objects' reward values. Second, to determine whether subjects' estimates were closer to estimates based on the feature-based or object-based approach, we computed the correlation between subjects' estimates and the actual reward probabilities assigned to each

object, or subjects' estimates and the reward probabilities estimated using average feature values (Eq.1).

Model fitting procedure. To capture subjects' learning and choice behavior, we used seven different reinforcement learning (RL) models based on object-based or feature-based approaches. These models were fit to experimental data by minimizing the negative log likelihood of the predicted choice probability given different model parameters using the 'fminsearch' function in MATLAB (Mathworks). We computed three measures of goodness-of-fit in order to determine the best model to account for the behavior in each experiment: average negative log likelihood, Akaike information criterion (AIC), and Bayesian information criterion (BIC). The smaller value for each measure indicates a better fit of choice behavior.

Object-based RL models. In this group of models, the reward value of each object is directly estimated from reward feedback on each trial using a standard RL model (Sutton & Barto, 1998). For example, in the uncoupled object-based RL, only the reward value of the chosen object is updated on each trial. This update is done via separate learning rates for rewarded or unrewarded trials using the following equations, respectively (Donahue & Lee, 2015):

$$V_{choO}(t+1) = V_{choO}(t) + \alpha_{rew}\big(1 - V_{choO}(t)\big)\,, \quad if\ R(t) = 1$$

$$V_{choO}(t+1) = V_{choO}(t) - \alpha_{unr}\big(V_{choO}(t)\big), \quad if\ R(t) = 0 \qquad \text{(Eq. 4)}$$

where *t* represents the trial number, V_{choO} is the estimated reward value of the chosen object, $R(t)$ is the trial outcome (1 for rewarded, 0 for unrewarded), and α_{rew} and α_{unr} are the learning rates for rewarded and unrewarded trials. The value of the unchosen object is not updated in this model.

In the coupled object-based RL, the reward values of both objects presented on a given trial are updated, but in opposite directions (assuming that reward assignments on the two objects are anti-correlated). That is, while the value of chosen object is updated based on Equation 4, the value of unchosen object is updated as follows:

$$V_{uncO}(t+1) = V_{uncO}(t) - \alpha_{rew}\big(V_{uncO}(t)\big), \quad if\ R(t) = 1$$

$$V_{unco}(t+1) = V_{unco}(t) + \alpha_{unr}\big(1 - V_{unco}(t)\big),\ \ if\ R(t) = 0 \qquad \text{(Eq. 5)}$$

where t represents the trial number and V_{unco} is the estimated reward value of the unchosen object.

The estimated value functions are then used to compute the probability of selecting between the two objects on a given trial (*O1* and *O2*) based on a logistic function

$$P_{O1}(t) = \frac{1}{1+\exp\left(-(V_{O1}(t)-V_{O2}(t))/\sigma\right)} \qquad \text{(Eq. 6)}$$

where P_{O1} is the probability of choosing object 1, V_{O1} and V_{O2} are the reward values of the two presented objects, and σ is a parameter measuring the level of stochasticity in decision process.

Feature-based RL models. In this group of models, the reward value (probability) of each object is computed using the average reward value of the features of that object, which are in turn estimated from reward feedback using a standard RL model. The update rules for the feature-based RL models are identical to the object-based ones, except that the reward value of the chosen (unchosen) object is replaced by the reward values of the features of the chosen (unchosen) object. In particular, only the reward values of unique features are updated on a given trial (e.g. if both objects are blue, then the reward value of blue is not updated). However, we also tested another feature-based RL where the reward values of both features are updated on each trial. The results from this model are not presented because the fit of this model was worse than that of other presented models and moreover, this model does not have an object-based counterpart.

As with the object-based RL models, the probability of choosing an object is determined based on the logistic function of the difference between the estimated values for the objects presented

$$P_{O1}(t) = \frac{1}{1+\exp\left(-\left(\left(V_{shapeO1}(t)-V_{shapeO2}(t)\right)+\left(V_{colorO1}(t)-V_{colorO2}(t)\right)\right)/2\sigma\right)} \qquad \text{(Eq. 7)}$$

where $V_{shapeO1}(V_{colorO1})$ and $V_{shapeO2}(V_{colorO2})$ are the reward values associated with the shape (color) of objects 1 and 2, respectively.

RL models with decay. Additionally, we investigated the effect of 'forgetting' the reward values of unchosen objects or features by introducing decay of value functions (in the uncoupled models) which has been shown to capture learning in a multi-dimensional task (Niv et al., 2015). More specifically, the reward values of unchosen objects or features decays to 0 with a rate of d, as follows:

$$V(t+1) = (1-d)V(t) \qquad \text{(Eq. 8)}$$

where t represents the trial number and V is the estimated reward probability of an object or a feature.

Computational models. To gain insights into the neural mechanisms underlying multi-dimensional decision-making, we examined two possible network models that could perform such a task (Fig.6A-B). Both models have two sets of value-encoding neurons that learned the reward values of individual objects (object-value-encoding neurons, OVE) or features (feature-value-encoding neurons, FVE). More specifically, plastic synapses onto value-encoding neurons undergo reward-dependent plasticity (via reward feedback), which enables these neurons to represent and update the value of presented objects or their features. Namely, reward values associated with individual objects and features are updated by potentiating or depressing plastic synapses onto neurons encoding the value of a chosen object or its features depending on whether the choice was rewarded or not rewarded, respectively.

The two network models differ in how they integrate signals from the OVE and FVE neurons and how the influence of signals from these neurons on the final choice is adjusted. The parallel decision-making and learning (PDML) model makes two additional decisions using the output of an individual set of value-encoding neurons (OVE or FVE) in order to compare with the choice of the final decision-making (DM) circuit (Fig.6A). If the final choice was rewarded, the model increases the strength of connections between the set or sets that produced the same choice as the final choice, therefore increasing the influence of the set of value-encoding neurons that were more likely responsible for making the final choice, and vice versa. By contrast, the hierarchical decision-making and learning (HDML) model updates connections from the OVE and FVE neurons to the corresponding neurons in the signal-selection circuit by determining which set of the value-encoding neurons contains a stronger signal (the difference between the values of the

two options) first, and uses only the outputs of that set to make the final decision on a given trial (Fig.6B). Subsequently, only the strengths of connections between the set of value-encoding neurons responsible for the 'selected' signal and the corresponding neurons in the signal-selection circuit are increased or decreased depending on whether the final choice was rewarded or not rewarded, respectively (see below).

Learning rule. We assumed that plastic synapses undergo a stochastic, reward-dependent plasticity rule (see Soltani and Wang, 2006 and Soltani, Lee, & Wang, 2006 for details). Briefly, we assumed that plastic synapses are binary and could be in potentiated (strong) or depressed (weak) states. On every trial, plastic synapses undergo stochastic modifications (potentiation or depression) depending on the model's choice and reward outcome (see below). During potentiation events, a fraction of weak synapses transition to the strong state with probability q_+. During depression events, a fraction of strong synapses transition to the weak state with probability q_-. These modifications allowed a given set of plastic synapses to estimate reward values associated with a choice option (Soltani, Lee, & Wang, 2006; Soltani & Wang, 2006, 2008, 2010).

For binary synapses, the fraction of plastic synapses that are in the strong state (which we call 'synaptic strength') determines the firing rate of afferent neurons. We denote the synaptic strength of plastic synapses onto a given population of value-encoding neurons 'v' by $F_v(t)$, where v = {R, B, s, t, Rs, Bs, Rt, Bt} represents a pool of neurons encoding the value of a given feature or a combination of features (in Experiments 1 and 2), and t represents the trial number. In Experiments 3 and 4, the number of feature values was three and four, respectively, instead of two, resulting in six and eight sets of FVE neurons and nine and sixteen sets of OVE neurons, respectively. Similarly, we denote the synaptic strength of plastic synapses from value-encoding neurons to the final DM circuit in the PDML model, or to the signal-selection circuit in the HDML model, by $C_m(t)$ where m = {O, F} represents general connections from OVE and FVE neurons, respectively.

The changes in the synaptic strengths for synapses onto value-encoding neurons depend on the model's choice and reward outcome on each trial. More specifically, we assumed that synapses

selective to the chosen object or features of the chosen object undergo potentiation or depression depending on whether the choice was rewarded or not, respectively:

$$F_{v(ch)}(t+1) = F_{v(ch)}(t) + q_{+}\left(1 - F_{v(ch)}(t)\right), \qquad if\ R(t) = 1$$

$$F_{v(ch)}(t+1) = F_{v(ch)}(t) - q_{-}F_{v(ch)}(t),\ if\ R(t) = 0 \quad \text{(Eq. 9)}$$

where t represents the trial number, $F_{v(ch)}(t)$ is the synaptic strength for synapses selective to the chosen object or features of the chosen object, $R(t)$ is the reward outcome, and q_{+} and q_{-} are potentiation and depression rates, respectively. The rest of plastic synapses transition to the weak state, according the following equation

$$F_{v(unch)}(t+1) = F_{v(unch)}(t) - q_d F_{v(unch)}(t) \qquad \text{(Eq. 10)}$$

where $F_{v(unch)}(t)$ is the synaptic strength for synapses selective to the unchosen object or features of the unchosen object, and q_d is the depression rate for the rest of plastic synapses.

We used similar learning rules for plastic synapses from value-encoding neurons to the final DM circuit in the PDML model as we did from value-encoding neurons to the signal-selection circuit in the HDML model. In the PDML model, plastic synapses from value-encoding neurons to the final DM circuit are updated depending on additional decisions based on the signal in an individual set of value-encoding neurons (OVE or FVE), the final choice, and the reward outcome as follows:

$$C_m(t+1) = C_m(t) + q_{+}\left(1 - C_m(t)\right),\ \ if\ R(t) = 1, and\ pool\ m\ choice = final\ choice$$

$$C_m(t+1) = C_m(t) - q_{-}C_m(t),\ \ if\ R(t) = 0, and\ pool\ m\ choice = final\ choice$$

$$C_m(t+1) = C_m(t)(1-d),\ \ if\ \ pool\ m\ choice \neq final\ choice \qquad \text{(Eq. 11)}$$

where t represents the trial number, $C_m(t)$ is the synaptic strength of connections from object-value-encoding ($m = O$) or feature-value-encoding neurons ($m = F$), q_{+} and q_{-} are potentiation and depression rates, respectively.

As we have shown before, the decision only depends on the overall difference in the output of the two value-encoding pools (Soltani, Lee, & Wang, 2006; Soltani & Wang, 2006, 2008, 2010). Importantly, if these pools are similar, this difference is proportional to the difference in the overall fraction of the strong synapses in the two pools (since we assumed binary values for synaptic efficacy). Therefore, the probability of the final choice in the PDML model depends on the difference between the sum of the output of the value-encoding neurons selective for the presented objects or their features (shape and color):

$$P(O_1) = \frac{1}{1+\exp\left(-\frac{C_O(F_{O1}-F_{O2})+C_F((F_{shapeO1}-F_{shapeO2})+(F_{colorO1}-F_{colorO2}))/2}{\sigma}\right)} \quad \text{(Eq. 12)}$$

where $F_{shapeOi}(t)$ and $F_{colorOi}(t)$ are the synaptic strengths for synapses onto FVE neurons selective to shape and color, respectively. The probabilities of additional decisions (in DM circuits 1 and 2) based on the signal in an individual set of value-encoding neurons (OVE or FVE) are computed by setting C_O or C_F in the above equation to zero.

In the HDML model, a signal-selection circuit determines which set of the value-encoding neurons (OVE or FVE) contains a stronger signal first, and uses only the output of that set to drive the final DM circuit on a given trial. The probability of selecting the signal from OVE neurons, $P(OVE)$, is computed using the following equation:

$$P(OVE) = \frac{1}{1+\exp\left(-\frac{C_O(F_{O1}-F_{O2})-C_F((F_{shapeO1}-F_{shapeO2})+(F_{colorO1}-F_{colorO2}))/2}{\sigma}\right)} \quad \text{(Eq. 13)}$$

Therefore, the final decision in the HDML model depends on the difference between the outputs of subpopulations in the set of value-encoding neurons which is selected as the set with stronger signal:

$$P(O_1) = \frac{1}{1+\exp\left(-\frac{F_{O1}-F_{O2}}{\sigma}\right)}, \ if\ OVE\ signal\ is\ selected$$

$$P(O_1) = \frac{1}{1+\exp\left(-\frac{F_{shapeO1}-F_{shapeO2}+F_{colorO1}-F_{colorO2}}{2\sigma}\right)}, \ if\ FVE\ signal\ is\ selected \quad \text{(Eq. 14)}$$

Finally, only plastic synapses from value-encoding neurons to the signal-selection circuit are updated depending on the final choice and the reward outcome:

$$C_m(t+1) = C_m(t) + q_+\big(1 - C_m(t)\big), \quad if\ R(t) = 1$$

$$C_m(t+1) = C_m(t)(1-d), \quad if\ R(t) = 0 \qquad \text{(Eq. 15)}$$

where m is the selected signal.

Model simulations. In order to test the behavior of the two network models during Experiments 1 and 2, we simulated each model over various environments with different levels of generalizability and volatility (Fig.7). More specifically, we linearly morphed a generalizable environment to a non-generalizable environment while modulating the level of volatility by changing the block length, L. For simulations of Experiments 3 and 4, we changed the levels of generalizability by randomly shuffling some of the elements of a fully generalizable matrix, using two sets of stimuli with different values of dimensionality (equal to n^m) while reward schedules remained fixed over the course of the experiment.

To assess the influence of non-generalizability, volatility, and dimensionality reduction on the behavior of each model, we used three measures. The first was performance, which was defined as the average harvested reward in a given environment. The second measure was the difference in connection strengths from value-encoding neurons to the final DM circuit in the PDML model or to signal-selection circuit in the HDML model. The connection strengths from the OVE/FVE neurons to the final DM circuit in the PDML model or signal-selection circuit in the HDML model were equated with the synaptic strength ($C_O(t)$ and $C_F(t)$) in the respective models. The third measure was the difference in overall weights that models assigned to object-based versus feature-based reward values. This measure combines the information in synapses onto value-encoding neurons with the strength of the output of these neurons (measure 2) and is computed as the product of the 'differential signals' (S) in a given set of value-encoding neurons and the synaptic strengths of connections from these neurons. The differential signal for the object-based reward values was computed by replacing $V_{O_i}(t)$ in Equation 3 with $F_{Oi}(t)$, which is the synaptic strength for synapses onto a pool i of OVE neurons. Similarly, the differential signal for the feature-based reward values was computed by using the estimated reward values for objects based on the synaptic strengths for synapses onto FVE neurons selective to shape and color ($F_{shape,i}(t)$ and $F_{color,i}(t)$) and Equation 1. Finally, the overall weight that the model assigned to

the object-based reward value (W_O(t)) was set equal to $C_O(t) \times S_O(t)$ and the overall weight assigned to the feature-based reward value (W_F(t)) was set equal to $C_F(t) \times S_F(t)$.

Models parameters. Both models have six parameters including potentiation and depression rates for plastic synapses onto value-encoding neurons ($q_+ = q_- = 0.15$), potentiation and depression rates for plastic synapses onto the final DM circuit in the PDML model or signal-selection circuit in the HDML model ($q_+ = q_- = 0.075$), the depression rate for the rest of plastic synapses ($q_d = 0.015$), and the level of stochasticity in choice and selection ($\sigma = 0.1$). Although we chose these specific parameter values for model simulations, the qualitative behavior of the models did not qualitatively depend on these parameters.

Results

Adaptability-precision tradeoff and feature-based learning.

We developed a general framework for understanding model adoption during learning reward values in multi-dimensional environments (Fig.3A). Assuming that objects (options) contain *m* features, each of which can take one of *n* feature values, there are n^m possible objects in the environment. The decision maker's task is to learn the reward values of options via reward feedback in order to maximize the total reward by choosing between two alternative options on each trial. In this framework, each object is assigned a reward value, although the average reward value for each value of a feature (e.g. red, green, triangles, squares) might be computed by averaging values of objects that contain that feature. By varying the relationship between the reward value of each option and the average reward values of its features, we generated multiple environments, each with a different level of generalizability (see Materials and Methods). To quantify generalizability, we defined the generalizability index as the correlation between reward values assigned to each object and reward values estimated based on feature values (Eq.1 in Materials and Methods). In a fully generalizable environment (generalizability index equal to 1), reward values of all options can be accurately constructed based on reward values of their features.

The object-based learning model directly learns the value of individual objects via reward feedback. In contrast, the feature-based model learns the average value of individual feature values by updating the values of all features of the object for which feedback was given, and uses these feature values to estimate the values of objects (Eq.1 in Materials and Methods). Clearly, with an unlimited amount of time, the object-based learning can perfectly learn all option values, whereas the accuracy of the feature-based model is limited by the generalizability of the environment. We designed a metric, referred to as the differential signal (see Materials and Methods), to quantify how well object-based or feature-based learning models can differentiate between options at a given point. By comparing the time course of the differential signal for the object-based and feature-based learning models using the same learning rate, we computed the time at which the object-based learning obtains more information than the feature-based learning model (the 'switch point'). A larger switch point indicates the superiority (better performance) of the feature-based learning relative to the object-based learning for a longer amount of time whereas zero switch point indicates that the object-based learning is always better (due to non-generalizability of the environment).

[Figure 3 about here]

The feature-based learning might be faster than the object-based learning because reward values of all features of the selected option are updated after each reward feedback in the feature-based learning model, whereas only the value of the selected option is updated in the object-based learning model. As a result, for sufficiently large values of generalizability (> 0.5), the feature-based learning model exhibits a stronger differential signal early on, but ultimately, the signal in the object-based learning model reaches that of the feature-based learning model (Fig.3A). Importantly, if the volatility of the environment increases and reward contingencies change more often than once per switch-point number of trials, feature-based learning is advantageous. As expected, the switch point increased as the learning rate decreased or as the generalizability increased (Fig.3A).

These simulation results demonstrate how the adaptability-precision tradeoff might favor the adoption of the feature-based over the model-based model in some environments. Because only the value of the selected option is updated after each reward feedback, object-based learning in a

volatile environment requires a higher learning rate, which comes at the cost of lower precision. Feature-based learning can mitigate this problem by speeding up the learning via more updates per feedback, instead of increasing the learning rate. However, feature-based learning loses its precision in a less generalizable environment more than it gains in precision due to multiple updates for each feedback, making it inferior to object-based learning (Fig.3A). Importantly, the advantage of feature-based over object-based leaning increases with the dimensionality of the environment, as the number of value updates per reward feedback increases with the number of features in each object (Fig.3B). Finally, the generalizability index, defined as the correlation between reward values assigned to each object and reward values estimated based on feature values (see Materials and Methods), by chance tends to assume larger values in an environment with greater dimensionality. This property further increases the advantage for feature-based learning in high-dimensional environments (Fig.3B inset).

Overall, our framework for learning reward values in a multi-dimensional environment illustrates that overcoming the adaptability-precision tradeoff is the main factor for adopting feature-based learning, and moreover, provides clear predictions about how different factors such as dimensionality reduction, generalizability, and volatility, might influence its adoption. More specifically, frequent changes in reward contingencies and high dimensionality should force the decision maker to adopt feature-based learning in order to reduce dimensionality and to increase adaptability without adding noise (Fig.3C). On the other hand, lack of generalizability of feature values to all object values should encourage them to adopt more accurate object-based learning, but feature-based learning should be still dominant in the beginning since it acquires reward information more quickly. We tested the influence of these factors in the following experiments.

Feature-based learning in dynamic generalizable environments.

To explore different factors that influence how humans adopt feature-based or object-based learning in dynamic, multi-dimensional environments, we designed a series of four experiments in which human subjects learned the reward values of different objects through reward feedback. In particular, we manipulated the relationship between the reward values of objects and their features (color, shape, etc.). In each trial of all the experiments, subjects chose between a pair of different objects that would yield reward with different probabilities.

In Experiment 1, a pairs of object in each trial was colored shapes and their reward probabilities unpredictably changed over time. Importantly, the feature-based and object-based approach required learning the same number of reward values (for 4 objects and their features, respectively) and thus, adopting the feature-based approach did not reduce dimensionality. Moreover, reward probabilities assigned to different objects could be estimated from values of their features (generalizable environment). By examining choice behavior during Experiment 1, we aimed to study specifically how adaptability required in a dynamic environment influences model adoption (Fig.3). Experiment 2 was similar to Experiment 1 except that reward probabilities assigned to different objects were not generalizable and could not be estimated from their feature values. Therefore, choice behavior in Experiment 2 could reveal how the adaptability required in a dynamic environment and a lack of generalizability influence model adoption (Fig.3). Finally, we increased the dimensionality to introduce a small (Experiment 3) and a moderate (Experiment 4) dimensionality reduction if feature-based learning was adopted, but fixed reward probabilities throughout the experiment. Reward values assigned to features, however, were not fully generalizable to objects, allowing us to study the influence of dimensionality reduction and lack of generalizability on model adoption (Fig.3).

During Experiments 1 and 2, subjects completed a two-alternative choice task (two sessions each with 768 trials) where on each trial they selected between two colored shapes (drawn from a set of four shapes; Fig.1A) that provided reward probabilistically. Importantly, reward probabilities assigned to individual objects (reward schedule) changed between blocks of trials (every 48 trials) in order to create environments with dynamic reward schedules. Overall, most subjects performed above the statistical chance level in both environments, indicating that they learned the values of options (Fig.4A-B). To examine the time course of learning, we computed the average probability of reward during each block of trials (i.e. when the reward probabilities were fixed). This analysis revealed that it took approximately 15 trials for the subjects to reach their maximum performance in each block (Fig.4C).

[Figure 4 about here]

To identify the learning model used by each subject, we fit the experimental data in each environment using various RL models that relied on either an object-based or a feature-based

approach (Materials and Methods). Two of the feature-based RLs, coupled feature-based RL and feature-based RL with decay, provided the best overall fit for the data in the generalizable environment (Experiment 1; Table 1). Both of these feature-based RLs provided a better fit than their corresponding object-based RLs, as measured by any of the goodness-of-fit indices (log likelihood, AIC, or BIC; Fig.4D-F). By contrast, in the non-generalizable environment (Experiment 2), the object-based RLs provided a significantly better fit than the feature-based RLs (Table 2). More specifically, two of the object-based RLs, coupled object-based RL and object-based RL with decay, provided a better fit for the data in the non-generalizable environment (Fig.4D-F). These results illustrate that subjects tend to adopt feature-based learning in the generalizable environment and object-based learning in the non-generalizable environment. Therefore, although a dynamic reward schedule encouraged subjects to use feature-based learning, which improves adaptability without compromising precision, a lack of generalizability led them to switch to slower but more precise object-based learning.

Feature-based learning in non-generalizable environments.

Our framework predicts that feature-based learning should be adopted initially until the acquired information derived from the object-based approach becomes comparable to information derived from the feature-based approach. To test this prediction, we designed two additional experiments (Experiments 3 and 4) in which human subjects learned the values of a larger set of objects in a static, non-generalizable environment (see Materials and Methods and Fig.2). Using a static environment, we aimed to isolate the influence of generalizability and dimensionality reduction on model adoption. Moreover, to assess the temporal dynamics of adopting feature-based and object-based approaches more directly, we asked subjects to provide their estimates of reward probabilities for individual objects during five or eight estimation blocks throughout the experiment. The reward assignment was such that one of the two features was partially informative about the reward value, while the other feature did not provide any information by itself (compare the average of values in individual columns or rows in Fig.2A-B).

Overall, the subjects were able to learn the task in Experiment 3, and the average performance (across all subjects) monotonically increased over time and plateaued at about 150 trials (Fig.5A). Examination of the estimated reward probabilities for individual objects also showed

an improvement over time, but more importantly, suggested a transition from a feature-based to an object-based approach as the experiment progressed. We utilized model fitting and correlation to identify the model adopted by the subjects from their reward probability estimates (see Materials and Methods). The fit of subjects' estimates revealed that the weight of the object-based approach, relative to that of the feature-based approach, was much smaller than 0.5 during the first estimation block but gradually increased over time (Fig.5B). In addition, the relative weight of bias (as an indication of subject's lack of discrimination between objects reward values) sharply dropped to a small value early in the experiment. Similarly, correlation analysis revealed that during early blocks, the estimates of only a small fraction of subjects were more correlated with actual reward probabilities than reward probabilities estimated using feature values, but this fraction increased over time (Fig.5C). The results of these two analyses illustrated that subjects initially adopted feature-based learning and gradually switched to object-based learning.

[Figure 5 about here]

We increased dimensionality of the environment in Experiment 4 more than in Experiment 3, in order to further examine the influence of dimensionality reduction on model adoption. The performance plateaued much earlier (approximately 75 trials) in Experiment 4, indicating faster learning than in Experiment 3 (Fig.5E). Moreover, the fit of subjects' estimates revealed that the relative weight of the object-based approach only slightly increased over time and plateaued at a small value, while the relative weight of bias plateaued at a value larger than the one in Experiment 3 (Fig.5F). Both of these results suggest stronger feature-based learning compared to object-based learning when dimensionality increased. Correlation analysis revealed a very similar pattern (Fig.5G).

We also fit the data from Experiments 3 and 4 using various RL models in order to identify the model used by the subjects (see Materials and Methods). We found that object-based RL with decay provided the best overall fit in Experiment 3 (Table 3). Importantly, this model provided a better fit than its corresponding feature-based RL. These results indicate that, overall, subjects adopted an object-based approach more often. Examination of the goodness-of-fit over time, however, showed that the object-based learning model provided a better fit later in the

experiment (Fig.5D). In contrast, feature-based RL with decay provided the best overall fit in Experiment 4 (Table 4). The fit of this model was better than the corresponding object-based learning model early in the experiment but later the fits of the two models became similar (Fig.5H), which is consistent with the results based on subjects' reward estimates. Nevertheless, the fit based on the object-based model never exceeded that of the feature-based model. During both Experiments 3 and 4, subjects transitioned from feature-based learning to object-based learning as feature values could not predict the reward values of all objects. However, an increase in dimensionality in Experiment 4 furthermore biased the behavior toward feature-based learning.

In summary, we found that in dynamic environments, human subjects adopted feature-based learning even when this approach does not reduce dimensionality. Subjects switched to learning individual option values (object-based learning) when feature values could not accurately predict all objects' values due to the lack of generalizable rules. Finally, in low-dimensional, static environments without generalizable rules, subjects still adopted feature-based learning first before gradually adopting object-based learning. Overall, these experimental results demonstrate that feature-based learning might be adopted mainly to improve adaptability without reducing precision. We next used network models in order to capture our experimental observations and to gain insights into neural mechanisms for model adoption during learning in dynamic environments.

Hierarchical decision-making and learning.

To understand neural mechanisms underlying model adoption in a multi-dimensional decision-making task, we examined two alternative network models that could perform such tasks (Fig.6A-B). Because of their architectures, we refer to these models as the parallel decision-making and learning (PDML) model and the hierarchical decision-making and learning (HDML) model. Both models have two sets of value-encoding neurons that learn the reward values of individual objects (object-value-encoding neurons, OVE) or features (feature-value-encoding neurons, FVE). The plastic synapses onto these value-encoding neurons undergo reward-dependent plasticity, enabling these neurons to represent and update the value of presented objects or their features at any given time (see Materials and Methods). Updating reward values

associated with individual objects and features is rather straightforward. In a given trial, plastic synapses onto neurons encoding the value of a chosen object or its features could be potentiated or depressed depending on whether the choice is rewarded or not rewarded, respectively, resulting in an increase or a decrease in reward values of those options/features. In contrast, there are many ways to integrate signals from the OVE and FVE neurons and adjust the influence of these neurons on the final choice.

The two network models are different in how this integration is done and how the influence of signals from the OVE and FVE neurons on the final decision is adjusted. The PDML model makes two additional decisions using the output of an individual set of value-encoding neurons (OVE or FVE) in order to compare with the choice of the final decision-making (DM) circuit (Fig.6A). If the final choice is rewarded, the model increases the strength of connections between the set (or sets) that produced the same choice as the final choice. This increases the influence of the set of value-encoding neurons that was more likely responsible for making the final choice. In contrast, if the final choice is not rewarded, the model decreases the strength of connections between the set (or sets) that produced the same choice as the final choice, decreasing the influence of the set of value-encoding neurons that was more likely responsible for making the final choice. The HDML model updates connections from the OVE and FVE neurons to their corresponding signal-selection circuits by determining which set of the value-encoding neurons contains a stronger signal (i.e. the difference between the values of the two options) first, and it then uses only the output of that set to make the final decision on a given trial (Fig.6B). Subsequently, only the strength of connections between the set producing the 'selected' signal and the corresponding neurons in the signal-selection circuit is increased or decreased depending on whether the final choice was rewarded or not rewarded, respectively (see Materials and Methods for more details).

[Figure 6 about here]

We used the two network models to simulate learning during our experiments. We first examined the behavior during Experiment 1 with a generalizable rule in a dynamic environment in which the reward probabilities were switched every 48 trials. The strength of connections from the OVE and FVE neurons to the final DM circuit in the PDML model or to the signal

selection circuit in the HDML model increased initially but at a much faster rate for FVE neurons (Fig.6C, D). This happened because on each trial both features of a selected object were updated and thus, synapses onto FVE neurons were updated twice as frequently as those onto OVE neurons. These faster updates enabled the FVE neurons to signal a correct response more often than the OVE neurons soon after a change in reward probabilities.

In the PDML model, the strength of connections between each of the value-encoding neurons and the final DM circuit represents how strongly those neurons drive the final DM circuit. Similarly, the strength of connections between each of the value-encoding neurons and the signal-selection circuit represents how strongly those neurons drive the final DM circuit in the HDML model. The overall influence of the object-based or feature-based approach, however, also depends on the signal encoded in plastic synapses onto the OVE and FVE neurons. Therefore, we used a combination of the signal represented in a given set of value-encoding neurons and the strength of connections between those neurons and the final DM circuit in the PDML model (or the signal-selection circuit in the HDML model) in order to compute the overall weight of an object-based or feature-based approach on the final choice (W_O and W_F, respectively; see Materials and Methods). We found that at the beginning of each block, (W_F - W_O) initially decreased but later increased in both models, indicating that both models assigned a larger weight to feature-based than object-based reward values, but this effect was greater in the HDML compared to the PDML model (Fig.6E).

To study how lack of generalizability and frequency of changes in reward contingencies (volatility) affect model adoption, we used the two network models to simulate various environments by changing the block length (number of trials where reward probabilities were fixed) and the level of generalizability (see Materials and Methods). The maximum and minimum levels of generalizability in these simulations correspond to environments used in Experiments 1 and 2, respectively. Both models were able to perform the task in various environments with different levels of volatility and generalizability, but the performance of the HDML model was higher in all environments (Fig.7A, D, G). More importantly, the difference in the strength of connection from FVE and OVE neurons ($C_F - C_O$) was more strongly modulated by generalizability and volatility in the HDML compared to the PDML model, indicating that HDML was better able to adjust the strength of connections from value-encoding

neurons (Fig.7B, E, H). As generalizability or volatility increased, connections between FVE neurons and the signal-selection circuit became stronger than connections between OVE neurons and the signal-selection circuit. Therefore, only the HDML model assigned larger weights to feature-based than object-based reward values (larger $W_F - W_O$) as the environment became more generalizable or volatile (Fig.7C, F, I). Overall, these results demonstrated that although both models were able to perform the task, the HDML model provided higher performance and exhibited stronger adjustment of connections from the value-encoding neurons to the next level of computation. Therefore, HDML was more successful in assigning proper weights to different types of learning according to reward statistics in the environment.

[Figure 7 about here]

Finally, we examined the interaction between dimensionality reduction and generalizability in adopting a model of the environment by simulating various environments in Experiments 3 and 4 using the two models. Because the dimensionality is a discrete number, we simulated choice behavior in two different environments with different numbers of objects (by having different number of feature values, $D = 3^2$ and $D = 4^2$) while changing the level of generalizability (see Materials and Methods). Consistent with simulation results for Experiments 1 and 2, an increase in generalizability caused both models to assign higher weights to feature-based than object-based reward values, but this effect was much stronger for the HDML model (larger positive slopes in Fig.8E-F compared with Fig.8B-C). An increase in dimensionality further biased both models to assign more weight to feature-based than object-based reward values. Overall, the simulation results for two alternative network models reveal that hierarchical decision-making and learning can provide a neural mechanism for adopting the model for learning in dynamic environments.

[Figure 8 about here]

Discussion

The framework proposed in this study for learning reward values in dynamic, multi-dimensional environments provides specific predictions about different factors that influence how humans adopt feature-based versus object-based learning to tackle the curse of dimensionality. Our

experimental results demonstrated that learning in dynamic environments tends to favor the feature-based approach because this approach not only reduces dimensionality but also improves adaptability without compromising precision. When precision is compromised due to non-generalizability of the rules assumed for feature-based learning, object-based learning is adopted more frequently. Importantly, feature-based learning is initially adopted, even in the presence of non-generalizable rules that only slightly reduce dimensionality and when reward contingencies do not change over time. This suggests that the main factor for adopting feature-based learning is to increase adaptability without compromising precision and therefore to overcome the adaptability-precision tradeoff (APT).

The APT sets an important constraint on learning about reward in a dynamic environment where reward values change over time. On solution to mitigate the APT is to adjust learning (e.g. the learning rate) over time (Khorsand et al., 2016, Society for Neuroscience abstract). Nevertheless, even with adjustable learning, the APT still persists and becomes more critical in multi-dimensional worlds, since the learner may never receive reward feedback on many unchosen options and feedback on chosen options is very limited. Importantly, adopting feature-based learning enables more updates after each reward feedback, which can greatly enhance the speed of learning without adding noise, similarly to other heuristic learning mechanisms (Jocham et al., 2016). Moreover, such learning allows estimation of reward values for options which have never been encountered before (Kahnt et al., 2012). Our results could explain why learning in young children (i.e. learning based on small number of samples) is dominated by attending to individual features (e.g. choosing a favorite color) to the extent that it can prevent them from performing in the dimension-switching task (Zelazo, Frye, & Rapus, 1996). Interestingly, this inability has been attributed to a failure to inhibit attention to the previously relevant (rewarding) feature (Kirkham, Cruess, & Diamond, 2003). Finally, by focusing on color and choosing a favorite color, children could evaluate all options based on their color and reduce the number of feature values to two (e.g. favorite and non-favorite), further reducing the dimensionality. Thus, our results explain that choosing a favorite color not only reduces dimensionality but also increases adaptation without compromising precision.

Even though rules used for feature-based learning are only partially generalizable in the real world, this non-generalizability may not prevent humans from using feature-based learning for

various reasons. First, simply due to chance, the generalizability (index) is larger for a higher dimensionality if there is at least one informative feature in the environment. Second, feature values related to different domains can be learned separately (color of fruits or color of cars), resulting in more generalizable rules in those domains. Finally, non-generalizability may never be detected due to a very large number of features and options.

In addition to mitigating the APT, feature-based learning is computationally more inexpensive and feasible than object-based learning, since it can be achieved using a small number of value-encoding neurons with pure feature selectivity. In contrast to recent theoretical work that has highlighted the advantage and importance of non-linear, mixed-selectivity representation for cognitive functions (Fusi, Miller, & Rigotti, 2016; Rigotti et al., 2013), our work points to the importance of pure feature selectivity for reward representation. Therefore, the advantage of mixed-selectivity representation could be specific to tasks with low dimensionality (in terms of reward structure) or when information does not change over time such as in object categorization tasks (Brincat & Connor, 2004; Gross, Rocha-Miranda, & Bender, 1972; Güçlü & van Gerven, 2014; Logothetis, Pauls, & Poggio, 1995). Our results suggest that learning about reward in dynamic environments might depend more strongly on value-encoding neurons with pure selectivity (i.e. neurons representing the reward value of an individual feature), since representations of such neurons can be adjusted more frequently over time due to more updates per feedback. Moreover, considering that neurons with pure feature selectivity are also crucial for saliency computations (Soltani & Koch, 2010), modulations of these neurons by reward could provide an effective mechanism for the modulation of attentional selection by reward as well (Khorsand, Moore, & Soltani, 2015).

Based on our results, we predict larger learning rates for neurons with highly mixed selectivity; otherwise, the information in these neurons would lag the information in pure feature-selective neurons. Moreover, we predict that the complexity of reward value representation is directly related to the stability of reward information in the environment. As the environment becomes more stable, learning the reward value of conjunctions of features and objects become more feasible and thus, more complex representation of reward values will emerge. These novel predictions could be tested in future experiments.

Selection between feature-based and object-based approaches is not all-or-none. As in our computational models, learning based on both approaches should occur simultaneously in two separate circuits, and arbitration between the two forms of learning might be required (Lee, Shimojo, & O'Doherty, 2014). Our modeling results show that such arbitration could happen via competition between two circuits based on the strength of signals in each circuit. Recently, Lee et al. (2014) have argued that arbitration between model-based and model-free learning might be accomplished based on the reliability of reward prediction error or state prediction error in the model-free and model-based systems, respectively. Nevertheless, whether these two neural systems are anatomically separate is still a topic of debate (Daw, Niv, & Dayan, 2005; Lee, Seo, & Jung, 2012; Shteingart & Loewenstein, 2014). The architecture of our proposed HDML model is in line with the idea that possibly distinct object-based and feature-based learning circuits could interact before a choice is made. It would be interesting to examine whether our proposed mechanisms for arbitration between object-based and feature-based circuits could be also applied for arbitration between model-based and model-free learning as well.

Despite the fact that naturalistic learning from reward feedback entails options with overlapping features, most studies of value-based learning and decision making in dynamic environments utilize one-dimensional objects where reward values are associated with choice alternatives based on the value of only one feature (e.g. color) (Behrens, Woolrich, Walton, & Rushworth, 2007; Donahue & Lee, 2015; Sugrue, Corrado, & Newsome, 2004). Only recently have some studies used multi-dimensional experimental paradigms to study learning from reward feedback and explored possible solutions for the curse of dimensionality (Eldar, Cohen, & Niv, 2013; Hunt, Dolan, & Behrens, 2014; Niv et al., 2015; Vaidya, 2015; Wilson & Niv, 2012; Wunderlich et al., 2011). These studies have suggested that multi-dimensional tasks might be solved in two general ways, either by constructing a simplified representation of the stimuli and learning only a small subset of features (Eldar et al., 2013; Niv et al., 2015), or by inferring the structure of the task and creating rules to estimate reward values of options based on their features without ignoring any features (model-based approach) (Wunderlich et al., 2011). Our results are more consistent with the idea that many dimensions can be considered at once using a model-based approach.

Our experimental results are qualitatively more compatible with a hierarchical decision-making and learning (HDML) model since the parallel decision-making and learning model does not show the sensitivity to experimental factors observed in human subjects. In the HDML model, the best sources of information were identified to make decisions and weights for the selected sources are successively updated according to reward feedback. Therefore, reward feedback alone can correctly adjust behavior toward a more object-based or a more feature-based approach, without any explicit optimization or knowledge of the environment. Interestingly, competition through stages of hierarchy has been suggested as an underlying mechanism behind multi-attribute decision making as well (Hunt et al., 2014; Jocham, Hunt, Near, & Behrens, 2012). The HDML model proposed in this study shares some components with the model of Hunt et al (2014), though it includes learning as well. Similarly, Wunderlich et al (2011) also suggested that the brain holds weights for all possible informative dimensions simultaneously, and these weights are updated on every trial.

In conclusion, we show that a tradeoff between adaptability and precision explains why humans adopt feature-based learning, especially in a dynamic environment. Because this type of learning is computationally inexpensive, our results suggest that neurons with pure selectivity could be crucial for learning in dynamic environments. Moreover, our work provides a missing framework for understanding how heterogeneity in reward representation emerges.

Acknowledgments

We would like to thank Deepak John and Suha Syed for help with earlier versions of the experiments. This work was supported by Neukom CompX grant to A.S. and NIH Grants (MH108629 and MH108643) to D.L.

References

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, *13*(4), 341–379.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Braun, D. A., Mehring, C., & Wolpert, D. M. (2010). Structure learning in action. *Behavioural Brain Research*, *206*(2), 157–165.

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, *7*(8), 880–886.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 473–492.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., & Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *The Journal of Neuroscience*, *33*(13), 5797–5805.

Donahue, C. H., & Lee, D. (2015). Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. *Nature Neuroscience*.

Eldar, E., Cohen, J. D., & Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, *16*(8), 1146–1153.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256.

Gross, C. G., Rocha-Miranda, C. E. de, & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. *Journal of Neurophysiology*, *35*(1), 96–111.

Güçlü, U., & van Gerven, M. A. (2014). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Brain's Ventral Visual Pathway. *arXiv Preprint arXiv:1411.6422*.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning: data mining, inference and prediction. *New York: Springer-Verlag*, *1*(8), 371–406.

Hunt, L. T., Dolan, R. J., & Behrens, T. E. (2014). Hierarchical competitions subserving multi-attribute choice. *Nature Neuroscience*, *17*(11), 1613–1622.

Jocham, G., Brodersen, K. H., Constantinescu, A. O., Kahn, M. C., Ianni, A. M., Walton, M. E., Behrens, T. E. (2016). Reward-guided learning with and without causal attribution. *Neuron*, *90*(1), 177–190.

Jocham, G., Hunt, L. T., Near, J., & Behrens, T. E. (2012). A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. *Nature Neuroscience*, *15*(7), 960–961.

Kahnt, T., Chang, L. J., Park, S. Q., Heinzle, J., & Haynes, J.-D. (2012). Connectivity-based parcellation of the human orbitofrontal cortex. *The Journal of Neuroscience*, *32*(18), 6240–6250.

Khorsand, P., Moore, T., & Soltani, A. (2015). Combined contributions of feedforward and feedback inputs to bottom-up attention. *Feedforward and Feedback Processes in Vision*, 86.

Kirkham, N. Z., Cruess, L., & Diamond, A. (2003). Helping children apply their knowledge to their behavior on a dimension-switching task. *Developmental Science*, *6*(5), 449–467.

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, *35*, 287–308. https://doi.org/10.1146/annurev-neuro-062111-150512

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687–699.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552–563.

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(4), 343–364.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience*, *35*(21), 8145–8157.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590.

Shteingart, H., & Loewenstein, Y. (2014). Reinforcement learning and human behavior. *Current Opinion in Neurobiology*, *25*, 93–98.

Soltani, A., & Koch, C. (2010). Visual saliency computations: mechanisms, constraints, and the effect of feedback. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(38), 12831–12843. https://doi.org/10.1523/JNEUROSCI.1517-10.2010

Soltani, A., Lee, D., & Wang, X.-J. (2006). Neural Mechanism for Stochastic Behavior During a Competitive Game. *Neural Networks*, *19*, 1075–1090.

Soltani, A., & Wang, X.-J. (2006). A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *The Journal of Neuroscience*, *26*(14), 3731–3744.

Soltani, A., & Wang, X.-J. (2008). From biophysics to cognition: reward-dependent adaptive choice behavior. *Curr Opini Neurobio*, *18*, 209–216.

Soltani, A., & Wang, X.-J. (2010). Synaptic computation underlying probabilistic inference. *Nature Neuroscience*, *13*(1), 112–9. https://doi.org/10.1038/nn.2450

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science (New York, N.Y.)*, *304*(5678), 1782–7. https://doi.org/10.1126/science.1094765

Sutton, R. S., & Barto, a G. (1998). Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*, *9*(5), 1054–1054. https://doi.org/10.1109/TNN.1998.712192

Vaidya, A. R. (2015). Neural Mechanisms for Undoing the "Curse of Dimensionality." *The Journal of Neuroscience*, *35*(35), 12083–12084.

Wilson, R. C., & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in Human Neuroscience*, *5*, 189.

Wunderlich, K., Beierholm, U. R., Bossaerts, P., & O'Doherty, J. P. (2011). The human prefrontal cortex mediates integration of potential causes behind observed outcomes. *Journal of Neurophysiology*, *106*(3), 1558–1569.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development*, *11*(1), 37–63.

Figure Legends

Figure 1. Timelines and reward schedules of Experiments 1 and 2. (**A**) On each trial, the subject chose between two objects (colored shapes) and was provided with reward feedback (reward or no reward) on the chosen object. The inset shows the set of all objects used during Experiments 1 and 2. (**B**) Alternative schedules for assigning reward probability to individual objects based on a generalizable rule (Experiment 1). Reward schedules are coded to show which feature (color or shape) is more informative and which feature values are more rewarding. For example, 'Rs' indicates that red objects are more rewarding than blue objects, squares are more rewarding than triangles, and color ('R') is more informative than shape ('s'). (**C**) Alternative schedules for assigning reward probability to individual objects based on a non-generalizable rule (Experiment 2). For these schedules, only one of the two features was on average informative about reward values (e.g. red for 'r1' schedule). (**D-E**) Examples of generalizable environments constructed by switching between blocks of generalizable reward schedules every 48 trials. (**F-G**) Examples of non-generalizable environments constructed by switching between blocks of or non-generalizable reward schedules every 48 trials.

Figure 2. Reward probabilities and objects used in Experiments 3 and 4. (**A-B**) During Experiment 3, reward probabilities were assigned to nine possible objects defined by combinations of two features (S, shape; P, pattern), each of which could take any of three values. Reward probabilities were assigned such that there was no generalizable rule based on feature values that could predict reward probabilities on all objects. Numbers in parentheses show the actual probability values used in the experiment due to limited resolution for reward assignment. For the set in A, the pattern was on average more informative about reward, whereas shape alone was not informative. The opposite was true about the set in B. Each subject performed the experiment twice, once when pattern was informative and once when shape was informative, using different sets of shapes and patterns. To shorten the experiment, we excluded object 'S3P3' from the choice set. (**C-D**). During Experiment 4, reward probabilities were assigned to sixteen possible objects defined by combinations of two features (S, shape; P, pattern), each of which could take any of four values. To shorten the experiment, we excluded objects with reward probability of 0.5 from the choice set. Conventions are the same as in A-B. (**E**) A sample estimation trial during Experiments 3 and 4. On each estimation trial, the subject estimated the

probability of reward on an individual object by pressing one of ten keys on the keyboard. (**F**) The set of possible shapes used in Experiments 3 and 4. For each session of the experiment, only three or four (for Experiments 3 or 4, respectively) of these shapes were used for a given subject (randomly chosen). (**G**) The set of possible patterns used in Experiments 3 and 4. For each session of the experiment, only three or four (for Experiments 3 or 4, respectively) of these patterns were used.

Figure 3. A framework for understanding model adoption during learning reward values in a dynamic, multi-dimensional environment. (**A**) Switch point is plotted as a function of the generalizability index of the environment for different values of the learning rate. The switch point increases with generalizability and slower learning rates, indicating that adaptability and precision influence model adoption. The arrow shows zero switch point indicating that the object-based learning is always superior. (**B**) Switch point is plotted as a function of generalizability separately for environments with different values of dimensionality. The advantage of feature-based over object-based learning increases with larger dimensionality. The inset shows the distribution of the generalizability index in each environment. (**C**) The object-based approach for learning multi-dimensional options requires learning n^m values, where there are m possible features and n values per feature in the environment, whereas the feature-based approach entails learning only $n*m$ values. A feature-based approach, however, is only useful if there are generalizable rules for estimating the reward values of options based on their feature values. A lack of generalizability should encourage using the more precise object-based approach. On the other hand, frequent changes in reward contingencies (dynamic environment) should increase the use of the faster, feature-based approach .

Figure 4. Dynamic reward schedules promote feature-based learning, whereas a lack of generalizability promotes object-based learning. (**A**) Performance (average reward) of subjects during Experiments 1 (generalizable environment) and 2 (non-generalizable environment). Dashed lines show the mean performance and solid lines show the threshold used for excluding subjects whose performance was not distinguishable from chance (0.5). (**B**) Time course of learning during each block of trials. Plotted is the average harvested reward on a given trial within a block across all subjects (the shaded areas indicate s.e.m.). The dashed line shows chance performance, and the solid lines show the maximum performance in the two

environments. (**C-E**) Comparison of different measures for goodness-of-fit, showing that subjects were more likely to adopt a feature-based approach in the generalizable environment and an object-based approach in the non-generalizable environment. Plotted are the three measures of the goodness-of-fit (-log likelihood (D), AIC (E), and BIC (F)) based on the feature-based and object-based RL with decay, separately for each environment. The insets show histograms of the difference in the goodness-of-fit indices from the two models for the generalizable (blue) and non-generalizable (red) environments. The dashed lines show the medians, and the star (double star) shows that the median is significantly different from zero at $p < 0.05$ ($p < 0.001$) using a two-tailed, sign-rank test.

Figure 5. Transition from the feature-based to object-based approach. (**A**) The time course of performance during Experiment 3. Shaded areas indicate s.e.m., and the dashed line shows chance performance, whereas the red and blue solid lines show the maximum performance using the feature-based and object-based approaches, respectively. Arrows mark the locations of estimation blocks throughout a session. For some subjects, there were only five estimation blocks indicated by black arrows. (**B**) The time course of model adoption measured by fitting subjects' estimates of reward probabilities. Plotted is the relative weight of object-based to feature-based approach, and the relative weight of bias over time. Dotted lines show the fit of data based on an exponential function. (**C**) The time course of model adoption measured via correlation. Plotted is the fraction of subjects for which the correlation between their reward estimates and actual reward probabilities was larger than the correlation between their reward estimates and probabilities estimated using the average feature values. The dotted line shows the fit of data based on an exponential function. (**D**) Transition from the feature-based to object-based approach revealed by the average goodness-of-fit over time. Plotted are the average negative log likelihood based on the feature-based model, object-based RL model, and the difference between object-based and feature-based models during Experiment 3. Shaded areas indicate s.e.m., and the dashed line shows the measure for chance prediction. (**E-H**) The same as in A-D, but during Experiment 4.

Figure 6. Architectures and performances of two alternative network models for multi-dimensional, decision-making tasks. (**A-B**). Architectures of the PDML (A) and the HDML (B) models (see Materials and Methods). (**C**) The time course of the overall strengths of plastic

synapses between OVE and FVE neurons and the final DM circuit (C_O and C_F) in the PDML model, or between OVE and FVE neurons and the signal-selection circuit (C_O and C_F) in the HDML model. These simulations were done for the generalizable environment (Experiment 1) where the block length was 48. (**D**) The difference between the C_F and C_O over time in the two models. (**E**) The overall weights of each set of value-encoding neurons on the final decision for the same set of simulations shown in panels C and D.

Figure 7. The effects of frequent changes in reward contingencies (volatility) and generalizability on the model behaviors. (**A**) Performance of the PDML model in various environments with different levels of volatility and generalizability. The color map shows the performance (average harvested reward) for a given value of block length (L) and the generalizability index. (**B**) The difference between the strengths of plastic synapses from FVE and OVE neurons onto the final DM circuit (C_F -C_O) in the PDML model. The color map shows ($C_F - C_O$) for a given value of block length (L) and the generalizability index. (**C**) The difference between the overall weights of FVE and OVE neurons on the final DM circuit ($W_F - W_O$) in the PDML model. (**D-F**) The same as in A-C but for the HDML model. (**G-I**) The difference between the performance, strengths of plastic synapses, and the overall weights in the HDML and PDML models.

Figure 8. The effects of dimensionality and generalizability on model behaviors. Simulations of the PDML and HDML models are shown in A-C and D-F, respectively. (**A, D**) Changes in performance as a function of the generalizability index, separately for two environments with 9 and 16 objects. The dotted red and blue curves show the maximal performance based on object-based (O) and feature-based (F) learning, respectively, for $D = 3^2$. The dashed curves show the results for $D = 4^2$. The gray and black arrows indicate the values of the generalizability index used in Experiments 3 and 4, respectively. (**B, E**) The difference between the strengths of plastic synapses from FVE and OVE neurons onto the final DM circuit ($C_F - C_O$) in the PDML model (B) and from FVE and OVE neurons onto the signal-selection circuit in the HDML model (E). (**C, F**) The difference between the overall weights of FVE and OVE neurons on the final DM circuit ($W_F - W_O$) in the PDML model (C), and from FVE and OVE neurons on the signal-selection circuit in the HDML model (F).

Table 1. Goodness-of-fit measures (-log likelihood, AIC, and BIC) averaged over all subjects (mean ± s.e.m.) for three feature-based RLs and their object-based counterparts for Experiment 1. The best model (feature-based RL with decay) and its object-based counterpart are highlighted in cyan and brown, respectively. Each feature-based RL was compared with its object-based counterpart using a two-tailed, sign-rank test. The significance level of the test is coded as: $0.05 < p < 0.1$ ($^{+}$), $0.01 < p < 0.05$ (*), $0.001 < p < 0.01$ (**), and $p < 0.001$ (***).

Table 2. Goodness-of-fit measures (-log likelihood, AIC, and BIC) for Experiment 2. The best model (object-based RL with decay) and its feature-based counterpart are highlighted in cyan and brown, respectively. Same format as in Table 1.

Table 3. Goodness-of-fit measures (-log likelihood, AIC, and BIC) for Experiment 3. The best model (object-based RL with decay) and its feature-based counterpart are highlighted in cyan and brown, respectively. Same format as in Table 1.

Table 4. Goodness-of-fit measures (-log likelihood, AIC, and BIC) for Experiment 4. The best model (feature-based RL with decay) and its object-based counterpart of the best model are highlighted in cyan and brown, respectively. Same format as in Table 1.

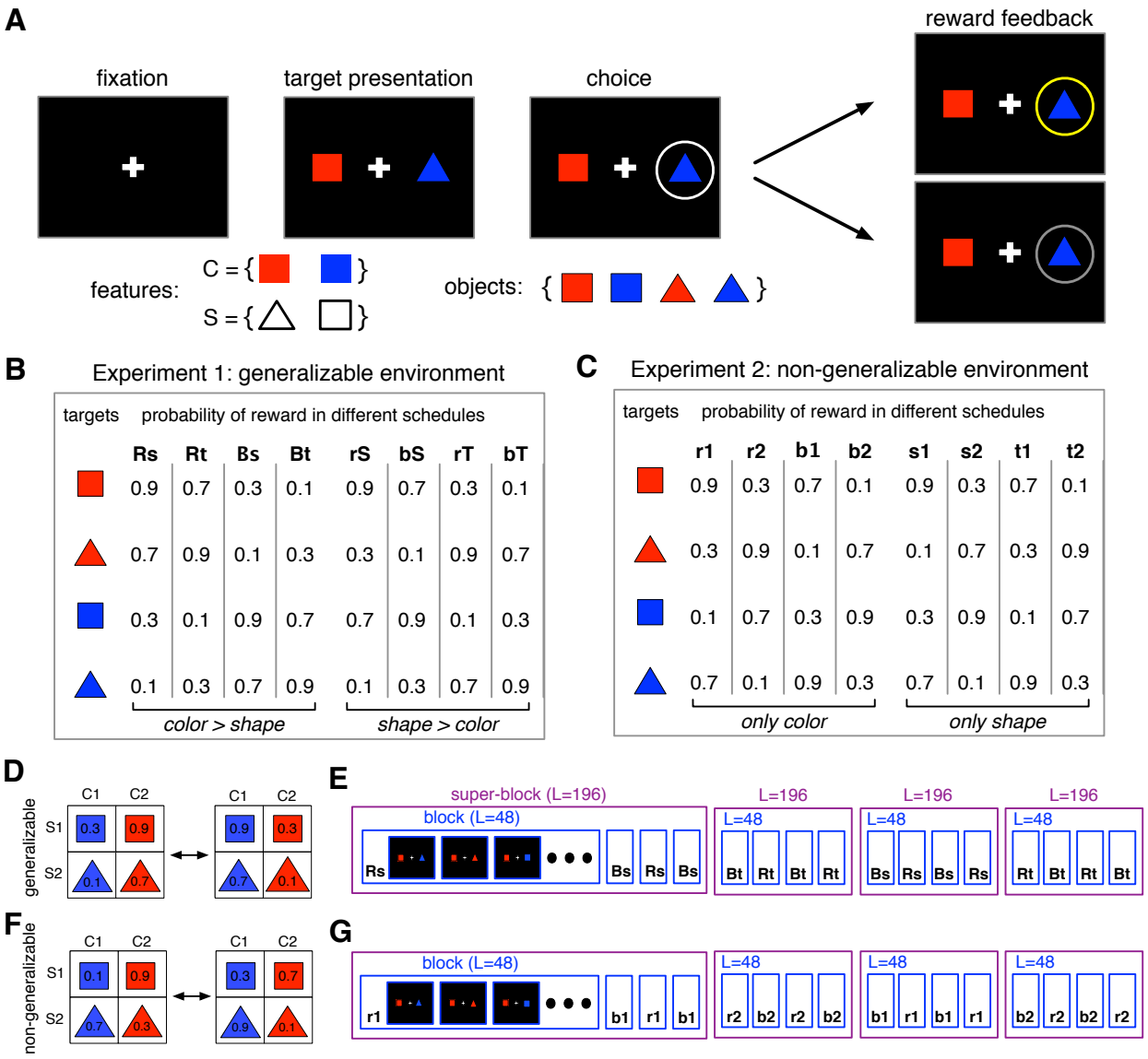
A
fixation
target presentation
choice
reward feedback
features:
C = { }
S = { }
objects: { }
B
Experiment 1: generalizable environment
targets
probability of reward in different schedules
Rs Rt Bs Bt rS bS rT bT
0.9 0.7 0.3 0.1 0.9 0.7 0.3 0.1
0.7 0.9 0.1 0.3 0.3 0.1 0.9 0.7
0.3 0.1 0.9 0.7 0.7 0.9 0.1 0.3
0.1 0.3 0.7 0.9 0.1 0.3 0.7 0.9
color > shape
shape > color
C
Experiment 2: non-generalizable environment
targets
probability of reward in different schedules
r1 r2 b1 b2 s1 s2 t1 t2
0.9 0.3 0.7 0.1 0.9 0.3 0.7 0.1
0.3 0.9 0.1 0.7 0.1 0.7 0.3 0.9
0.1 0.7 0.3 0.9 0.3 0.9 0.1 0.7
0.7 0.1 0.9 0.3 0.7 0.1 0.9 0.3
only color
only shape
D
generalizable
C1 C2
S1
S2
0.3 0.9 0.1 0.7
0.9 0.3 0.7 0.1
E
super-block (L=196)
block (L=48)
L=196
L=48
Rs Bs Rs Bs
Bt Rt Bt Rt
Bs Rs Bs Rs
Rt Bt Rt Bt
F
non-generalizable
C1 C2
S1
S2
0.1 0.9 0.7 0.3
0.3 0.7 0.9 0.1
G
block (L=48)
L=48
r1 b1 r1 b1
r2 b2 r2 b2
b1 r1 b1 r1
b2 r2 b2 r2

Figure 1

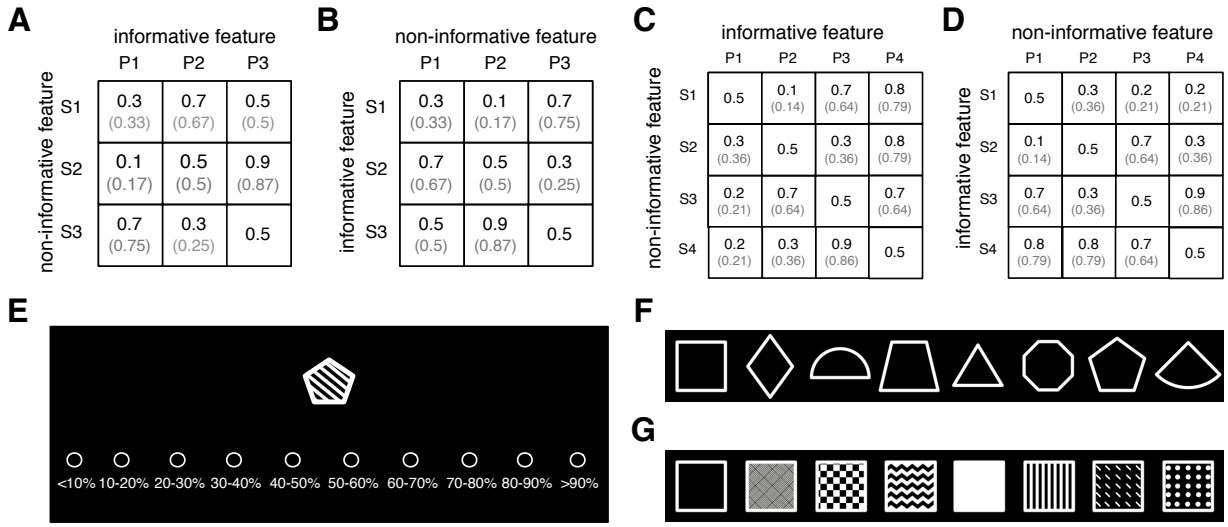
A
informative feature
P1 P2 P3
non-informative feature
S1 0.3 (0.33) 0.7 (0.67) 0.5 (0.5)
S2 0.1 (0.17) 0.5 (0.5) 0.9 (0.87)
S3 0.7 (0.75) 0.3 (0.25) 0.5
B
non-informative feature
P1 P2 P3
informative feature
S1 0.3 (0.33) 0.1 (0.17) 0.7 (0.75)
S2 0.7 (0.67) 0.5 (0.5) 0.3 (0.25)
S3 0.5 (0.5) 0.9 (0.87) 0.5
C
informative feature
P1 P2 P3 P4
non-informative feature
S1 0.5 0.1 (0.14) 0.7 (0.64) 0.8 (0.79)
S2 0.3 (0.36) 0.5 0.3 (0.36) 0.8 (0.79)
S3 0.2 (0.21) 0.7 (0.64) 0.5 0.7 (0.64)
S4 0.2 (0.21) 0.3 (0.36) 0.9 (0.86) 0.5
D
non-informative feature
P1 P2 P3 P4
informative feature
S1 0.5 0.3 (0.36) 0.2 (0.21) 0.2 (0.21)
S2 0.1 (0.14) 0.5 0.7 (0.64) 0.3 (0.36)
S3 0.7 (0.64) 0.3 (0.36) 0.5 0.9 (0.86)
S4 0.8 (0.79) 0.8 (0.79) 0.7 (0.64) 0.5
E
<10% 10-20% 20-30% 30-40% 40-50% 50-60% 60-70% 70-80% 80-90% >90%
F
G

Figure 2

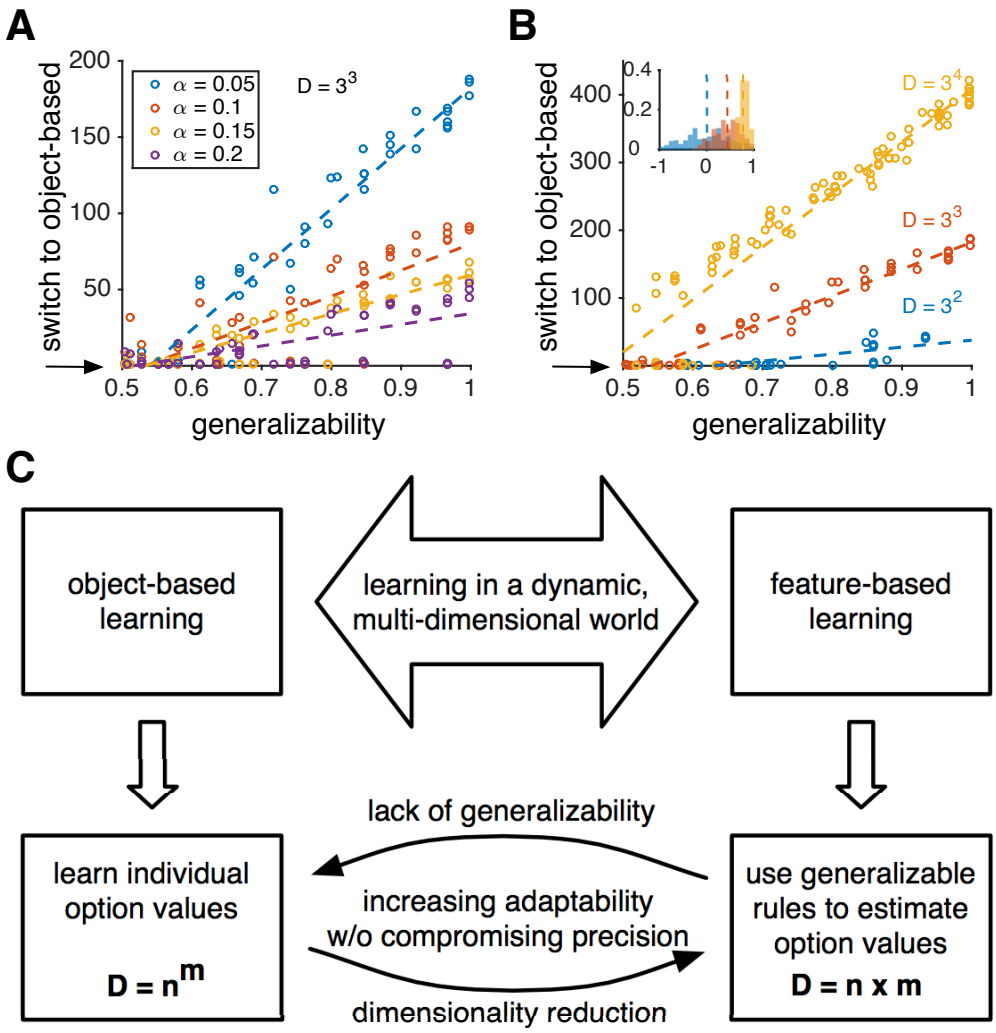
A
B
C
switch to object-based
generalizability
α = 0.05
α = 0.1
α = 0.15
α = 0.2
D = 3^3
D = 3^4
D = 3^3
D = 3^2
object-based learning
learning in a dynamic, multi-dimensional world
feature-based learning
lack of generalizability
learn individual option values
D = n^m
increasing adaptability w/o compromising precision
use generalizable rules to estimate option values
D = n x m
dimensionality reduction

Figure 3

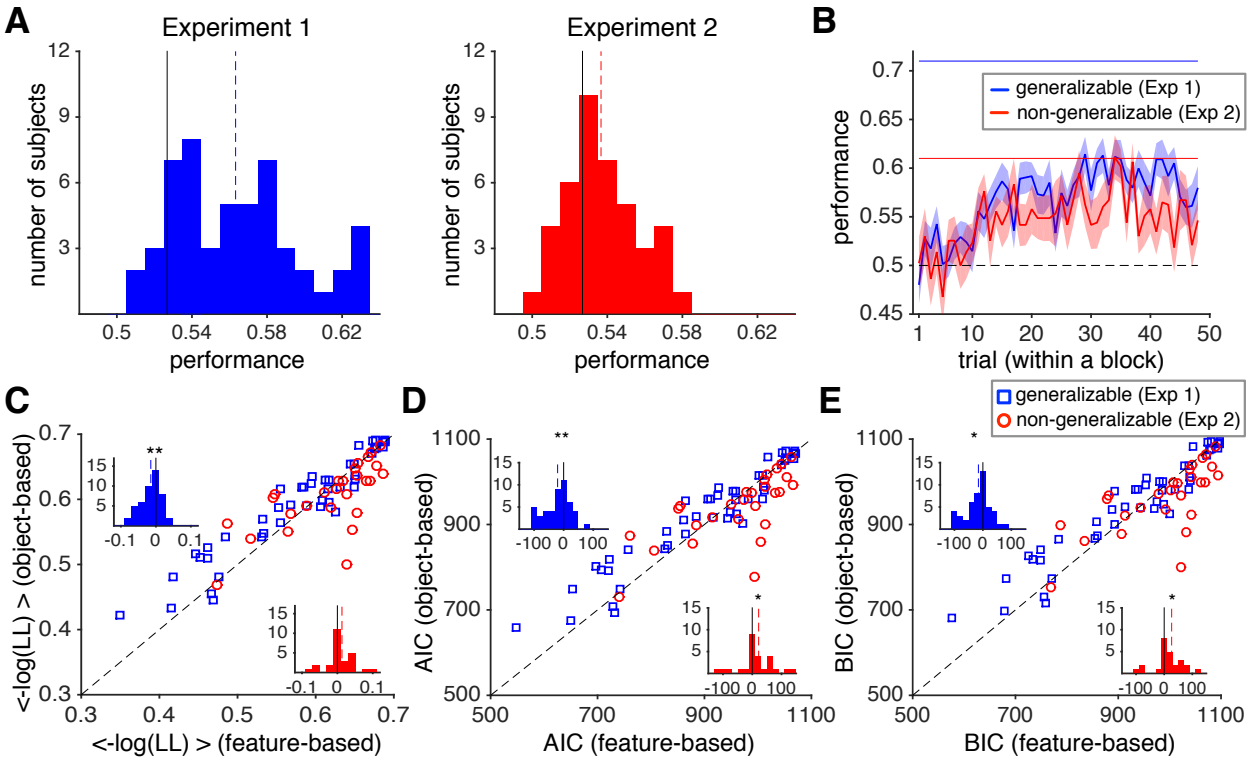
A
Experiment 1
Experiment 2
number of subjects
performance
B
generalizable (Exp 1)
non-generalizable (Exp 2)
performance
trial (within a block)
C
D
E
<-log(LL) > (object-based)
<-log(LL) > (feature-based)
AIC (object-based)
AIC (feature-based)
BIC (object-based)
BIC (feature-based)

Figure 4

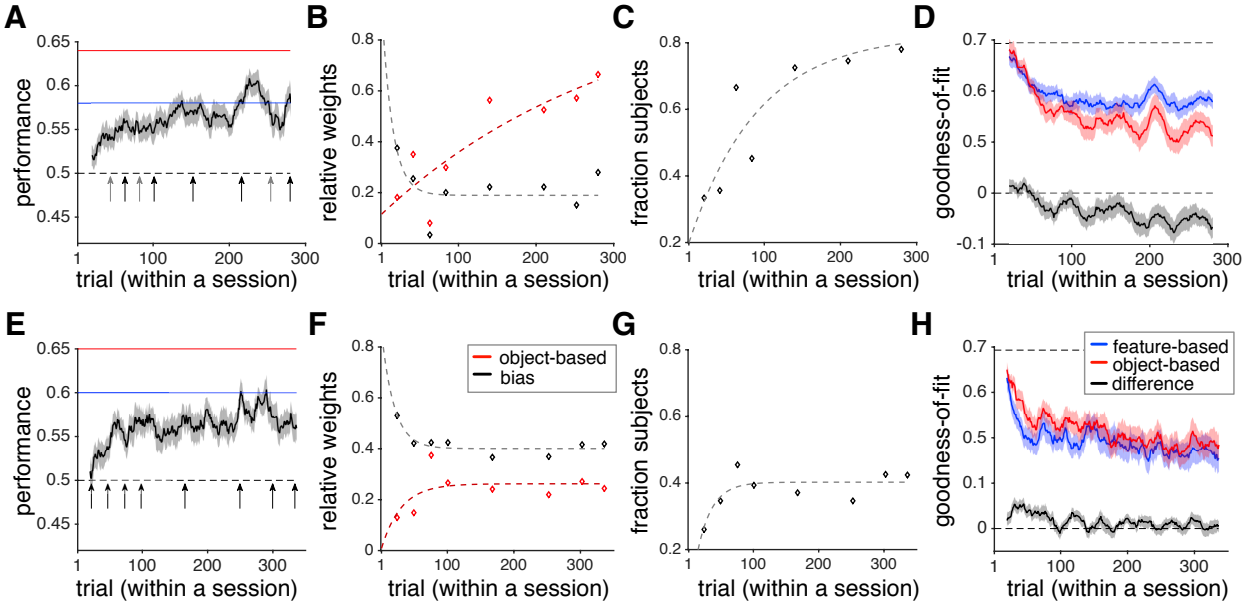
A
B
C
D
E
F
G
H
performance
relative weights
fraction subjects
goodness-of-fit
trial (within a session)
object-based
bias
feature-based
object-based
difference

Figure 5

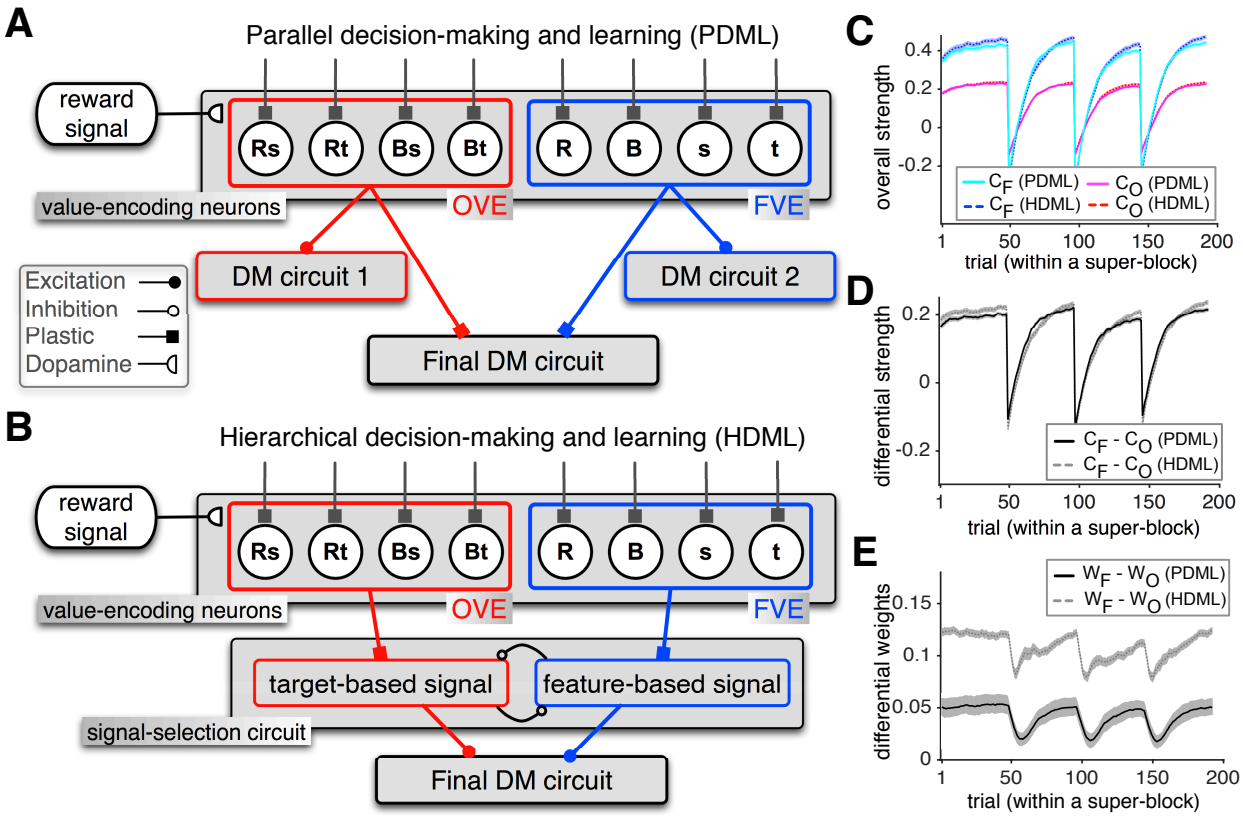
A
Parallel decision-making and learning (PDML)
reward signal
Rs
Rt
Bs
Bt
R
B
s
t
value-encoding neurons
OVE
FVE
Excitation
Inhibition
Plastic
Dopamine
DM circuit 1
DM circuit 2
Final DM circuit
B
Hierarchical decision-making and learning (HDML)
reward signal
Rs
Rt
Bs
Bt
R
B
s
t
value-encoding neurons
OVE
FVE
target-based signal
feature-based signal
signal-selection circuit
Final DM circuit
C
overall strength
0.4
0.2
0
-0.2
C_F (PDML)
C_O (PDML)
C_F (HDML)
C_O (HDML)
1
50
100
150
200
trial (within a super-block)
D
differential strength
0.2
0
-0.2
C_F - C_O (PDML)
C_F - C_O (HDML)
1
50
100
150
200
trial (within a super-block)
E
differential weights
0.15
0.1
0.05
0
W_F - W_O (PDML)
W_F - W_O (HDML)
1
50
100
150
200
trial (within a super-block)

Figure 6

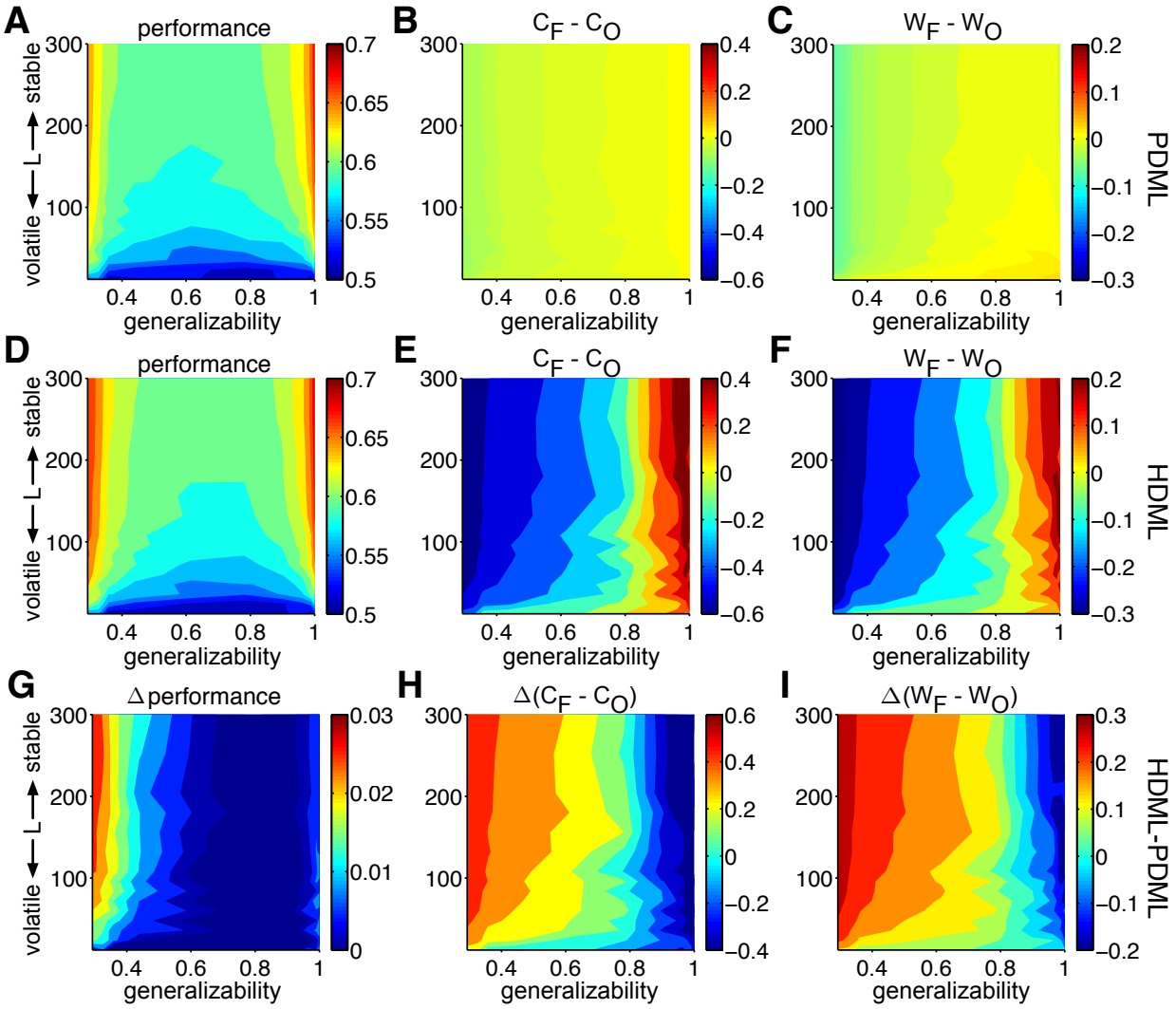
A
performance
B
C_F - C_O
C
W_F - W_O
PDML
D
performance
E
C_F - C_O
F
W_F - W_O
HDML
G
Δ performance
H
Δ(C_F - C_O)
I
Δ(W_F - W_O)
HDML-PDML
volatile ← L → stable
generalizability

Figure 7

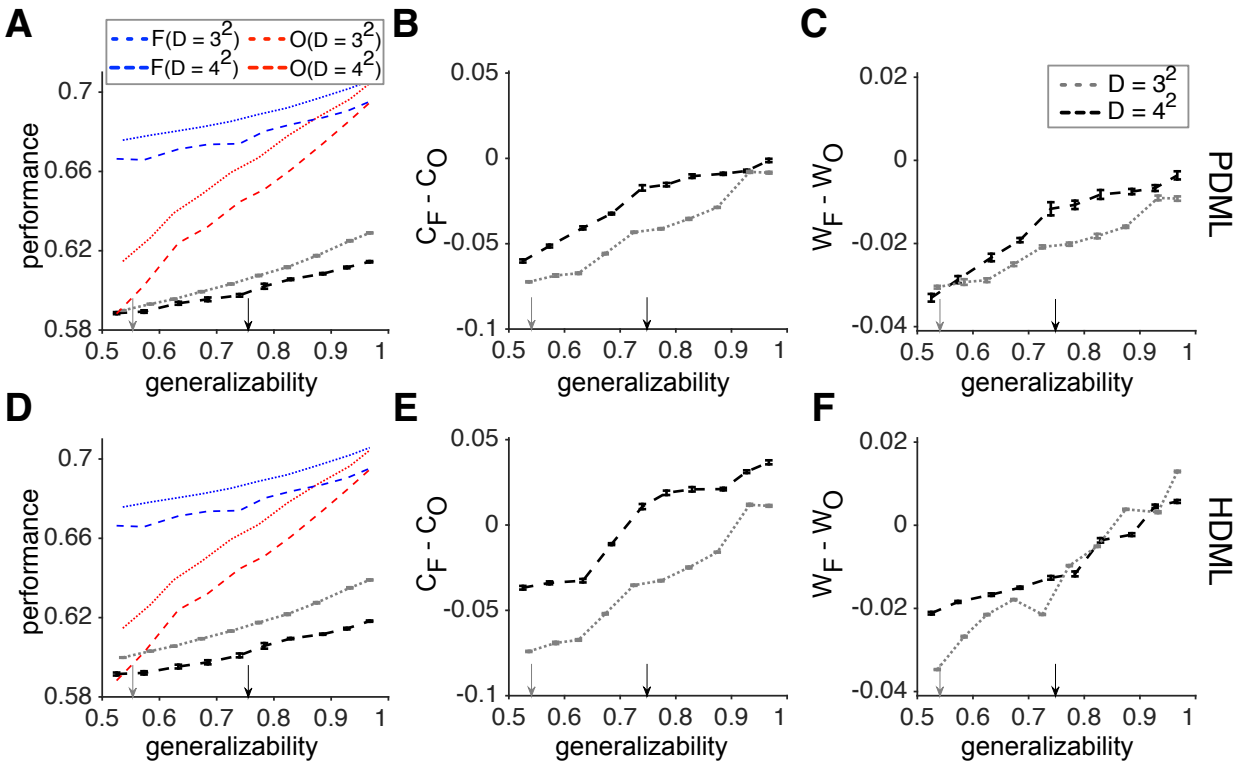
A
B
C
D
E
F
F(D = 3^2)
O(D = 3^2)
F(D = 4^2)
O(D = 4^2)
D = 3^2
D = 4^2
PDML
HDML
performance
generalizability
c_F - c_O
w_F - w_O
0.7
0.66
0.62
0.58
0.05
0
-0.05
-0.1
0.02
-0.02
-0.04
0.5
0.6
0.7
0.8
0.9
1

Figure 8

Table 1. Goodness-of-fit measures for Experiment 1.

Model	Coupled feature-based RL	Uncoupled feature-based RL	Feature-based RL with decay	Coupled object-based RL	Uncoupled object-based RL	Object-based RL with decay
Number of parameters	5	5	6	4	4	5
-log likelihood	443.6±10.9***	463.8±8.5***	446.4±9.9***	462.9±8.2	519.1±2.2	457.6±8.4
AIC	897.3±21.8***	937.5±17.0***	904.8±19.8***	933.7±16.5	1046.3 ±4.5	925.2±16.8
BIC	920.5±21.8***	960.7±17.0***	932.7±19.8*	952.3±16.4	1064.8±4.5	948.4±16.8

Table 2. Goodness-of-fit measures for Experiment 2.

Model	Coupled feature-based RL	Uncoupled feature-based RL	Feature-based RL with decay	Coupled object-based RL	Uncoupled object-based RL	Object-based RL with decay
Number of parameters	5	5	6	4	4	5
-log likelihood	479.2±8.0	491.6±6.9	474.1±8.7	469.1±8.1	512.0±6.2**	464.2±7.4^{+}
AIC	968.5±16.0	993.3±13.8	946.1±17.4	946.1±16.2^{+}	1032.1±12.4**	938.4±14.8*
BIC	991.7±16.0	1016.5±13.8	988.1±17.4	964.7±16.2*	1050.7±12.4**	961.6±14.9*

Table 3. Goodness-of-fit measures for Experiment 3.

Model	Coupled feature-based RL	Uncoupled feature-based RL	Feature-based RL with decay	Coupled object-based RL	Uncoupled object-based RL	Object-based RL with decay
Number of parameters	5	5	6	4	4	5
-log likelihood	352.9±4.8***	342.9±4.2	330.9±4.6	369.2±4.3	346.4±5.4	313.2±7.5*
AIC	717.5±9.7***	715.1±8.4	677.5±9.3	748.6±8.7	713.0±10.8	633.3±15.1**
BIC	747.8±9.7***	745.4±8.4	712.2±9.3	766.0±8.7	730.3±10.8	655.1±15.1***

Table 4. Goodness-of-fit measures for Experiment 4.

Model	Coupled feature-based RL	Uncoupled feature-based RL	Feature-based RL with decay	Coupled object-based RL	Uncoupled object-based RL	Object-based RL with decay
Number of parameters	5	5	6	4	4	5
-log likelihood	411.3±5.8***	387.2±5.5***	340.1±7.6**	458.4±1.7	429.2±5	354.9±5.6
AIC	832.7±11.7***	784.4±11***	692.1±15.2*	924.8±3.4	866.5±10	719.7±11.2
BIC	855.2±11.7***	807±11***	719.2±15.2^{+}	942.8±3.4	884.6±10	742.3±11.2