Your favorite color makes learning more adaptable and precise

Shiva Farashahi[1]*, Katherine Rowe[1]*, Zohra Aslami[1], Daeyeol Lee[2-5], Alireza Soltani[1]

[1]*Department of Psychological and Brain Sciences, Dartmouth College, NH, 03784*
[2]*Department of Neuroscience, Yale School of Medicine, New Haven, CT, 06510*
[3]*Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT, 06510*
[4]*Department of Psychiatry, Yale School of Medicine, New Haven, CT 06511*
[5]*Department of Psychology, Yale University, New Haven, CT, 06520*

* *These authors contributed equally to this work*

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

Manuscript information: 51 pages; 7 figures, 1 table, and 5 supplementary figures.
Abstract 150 words
Introduction 840 words
Results 5600 words (including figure captions)
Discussion 1550 words

April 19, 2017

Acknowledgments: We would like to thank Deepak John and Suha Syed for help with earlier versions of the experiments. This work was supported by National Science Foundation (EPSCoR Award #1632738) grant to AS, and National Institute of Health grants (MH108629 and MH108643) to DL.

Abstract

Learning from reward feedback is essential for survival but can become extremely challenging with myriad varying choice options. Here, we propose that learning reward values of individual features can provide a heuristic for estimating reward values of choice options in dynamic, multi-dimensional environments. We hypothesized 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 hypothesis 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 environments, subjects initially adopted feature-based learning and gradually switched to learning reward values of individual options, depending on how accurately objects' values can be predicted by combining feature values. Our computational models reproduced these results and highlighted the importance of neurons coding feature values for parallel learning of values for features and objects.

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 but can be extremely challenging in natural settings 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 must be learned. This is referred to as the "curse of dimensionality," because the standard reinforcement learning (RL) models used to simulate human learning do not scale up with the increasing dimensionality and number of possible options in the environment (Barto & Mahadevan, 2003; Botvinick, 2012; Hastie, Tibshirani, & Friedman, 2001; Sutton & Barto, 1998).

An increase in dimensionality creates two main difficulties for humans and the standard RL models that try to directly learn the value of individual options. First, learning is too slow and imprecise due to the large amount of reward feedback needed for an accurate estimate of the reward value, while reward contingencies might quickly change over time. For example, a child naturally learns the tastes of various fruits she consumes throughout her life (e.g. green crispy apples, red crispy apples, yellow mushy bananas, etc.), but it would take a long time to acquire preferences for all different types of fruits. Second, the value of unexperienced options cannot be known; for example, how should the child approach a green, mushy avocado never encountered before?

A few approaches are proposed for how we 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). There are behavioral and neural data, however, suggesting that in order to make decisions in multi-dimensional tasks, humans process all features of each option simultaneously, rather than focusing on a single feature at a time (Wunderlich, Beierholm, Bossaerts, & O'Doherty, 2011). Moreover, ignoring certain features could be detrimental in dynamic environments where previously non-informative features can suddenly become informative. Another approach is to combine multiple environmental states or actions, thereby reducing the number of states or actions to be learned (Botvinick, 2012; Ribas-Fernandes, et al., 2011). Finally, one could 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 reward values of all features are learned in parallel, and then combined according to a specific rule to estimate the reward values for individual options. For example, a child could evaluate fruits based on their color and texture and learn about these features when she consumes them. This heuristic feature-based learning, however, is only beneficial if a generalizable set of rules exist such that the reward value of all options can be closely constructed by combining the reward values of their features. Unfortunately, this is often not the case; for example, not all green fruits are tasty. So could the benefits of feature-based learning overcome a lack of generalizable rules in order to make this learning a viable heuristic?

An important aspect of feature-based learning is that reward values of all features of the selected option can be updated based on a single reward feedback, instead of only the value of the selected option in object-based learning. This makes feature-based learning faster and more adaptable, without being noisier, than object-based learning. This is important because simply increasing the learning rates in object-based learning can improve adaptability but also adds noise in the estimation of reward values, which we refer to as the adaptability-precision tradeoff (Farashahi, et al., 2017). Therefore, the main advantage of heuristic feature-based learning might be to overcome the adaptability-precision tradeoff. To test this hypothesis, we constructed a general framework for understanding the advantages of this learning and designed a series of experiments to characterize how multiple factors encourage the adoption of feature-based versus object-based learning. Moreover, we designed and tested two alternative network models to elucidate neural mechanisms consistent with our experimental observations.

We found that in dynamic environments, humans adopted feature-based learning even when this approach did not reduce dimensionality. Even in static, low-dimensional environments where dimensionality reduction was small, subjects initially adopted feature-based learning and only gradually switched to learning individual option/object values. The degree of switching to

object-based learning, however, was smaller with higher dimensionality or when objects' values could be more accurately predicted by combining the reward values of their features (i.e. more generalizable environment). Overall, these results confirmed our hypothesis and suggest feature-based learning as a powerful heuristic for learning in dynamic, multi-dimensional environments. Finally, we found that hierarchical decision-making and learning processes can account for our experimental data. These results highlight the importance of parallel learning of the reward values associated with features and objects.

Results

Feature-based learning could mitigate the adaptability-precision tradeoff. To test our hypothesis that feature-based learning is mainly adopted to overcome the adaptability-precision tradeoff, we first developed a general framework for learning in dynamic, multi-dimensional environments (see Methods for more details). If options/objects contain m features, each of which can have n types, there would be n^m possible objects in the environment. The decision maker's task is to learn the reward values of options/objects via reward feedback in order to maximize the total reward when choosing between two alternative options on each trial. To examine the advantages of object-based and feature-based approaches, we simulated this task using two different model learners. The object-based learner directly estimates the reward values of individual objects via reward feedback, whereas the feature-based learner estimates the reward values of all feature instances, such as red, blue, square, or triangle. The latter is achieved by updating the reward values associated with all features of the object for which reward feedback is given. The feature-based learner then combines the reward values of features to estimate the reward values of individual objects. To examine how the performance of the two learners depend on the reward statistics in the environment, we varied the relationship between the reward value of each object and the reward values of its features in order to generate multiple environments, each with a different level of generalizability (see Methods). In a fully generalizable environment, the estimated reward probabilities based of features deviate from the actual reward probabilities only by a small amount (Supplementary Figure 1), but more importantly, the order of estimated and actual probabilities is similar.

Feature-based learning might be faster than object-based learning when using the same learning rate, because reward values of all features of the selected option can be updated after each reward feedback in the feature-based learning model. In contrast, only the value of the selected option is updated in the object-based learning model. Clearly, given a sufficient amount of time, the object-based learner can accurately estimate all option values, whereas the accuracy of the feature-based learner is limited by the generalizability of the environment. By comparing the time course of information acquired by the object-based and feature-based learners when reward values are fixed and using the same learning rate, we computed the time at which the object-based learner obtains more information than the feature-based learner (the 'cross-over point'; see Methods).

We found that for sufficiently large values of generalizability (> 0.5), the feature-based learner acquires more information early on, but ultimately, the object-based learner reaches the same level of information as the feature-based learner and later surpasses it. This means that in a stable environment, object-based learning will be ultimately more useful. However, in volatile environments where reward contingencies change often, feature-based learning might be more beneficial. Moreover, the cross-over point happens later for smaller learning rates, indicating that slowing down learning to increase precision would favor feature-based learning (Fig. 1a). The advantage of feature-based over object-based learning 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. 1b). Finally, an environment with randomly assigned reward probabilities tends to be more generalizable as the dimensionality increases (Fig. 1b inset). This property further increases the advantage of adopting feature-based learning in high-dimensional environments.

These simulations demonstrate how the presence of the adaptability-precision tradeoff might favor the adoption of feature-based over object-based learning 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.

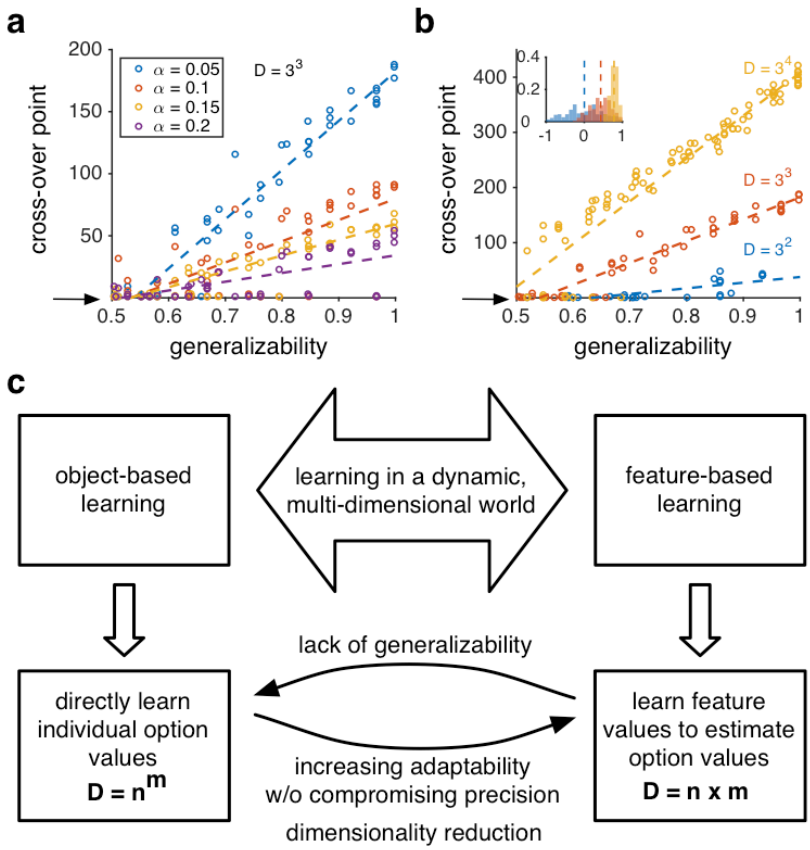


Figure 1. A framework for understanding model adoption during learning in dynamic, multi-dimensional environments. (**a**) Cross-over point is plotted as a function of the generalizability index of the environment for different values of the learning rate. The cross-over point increases with generalizability and decreases with the learning rate. The larger learning rate, however, comes at the cost of more noise in estimation (lower precision). The arrow shows zero cross-over point indicating that the object-based learning is always superior for certain environments. (**b**) Cross-over 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 randomly generated environments for three different dimensionalities. (**c**) The object-based approach for learning multi-dimensional options/objects requires learning n^m values, where there are m possible features and n types per feature in the environment, whereas the feature-based approach entails learning only $n*m$ values. A feature-based approach, however, is beneficial if there are generalizable rules for estimating the reward values of options based on the combination of features' values. A lack of generalizability should encourage using the object-based approach. On the other hand, frequent changes in reward contingencies (dynamic environment) should increase the use of feature-based learning, which can increase adaptability without compromising precision.

Our simple framework also provides clear predictions about how different factors such as dimensionality reduction, generalizability, and volatility might influence the adoption of feature-based learning. 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. 1c). On the other hand, lack of generalizability of the reward values of features to all object values should encourage adopting more accurate object-based learning, but immediately after changes in reward values, feature-based learning should still be favored since it acquires reward information more quickly. We tested the influence of these factors in four experiments.

Feature-based learning in dynamic environments. To test our hypothesis and explore different factors that influence how humans adopt feature-based versus object-based learning in dynamic multi-dimensional environments, we designed 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 those of their features (color, shape, etc.). In each trial, subjects chose between a pair of dissimilar objects associated with different reward probabilities.

In Experiment 1, the pair of objects in each trial consisted of colored shapes whose reward probabilities unpredictably changed over time. Importantly, the feature-based and object-based approach required learning the same number of reward values: four objects (red square, red triangle, blue square, and blue triangle) and four feature instances (red, blue, square, and triangle). Therefore, adopting feature-based learning did not reduce dimensionality in Experiment 1. Moreover, reward probabilities assigned to different objects could be closely estimated by combining the reward values of their features; that is, the environment was generalizable (Supplementary Figure 1). By examining choice behavior during Experiment 1, we aimed to study specifically how adaptability required in a dynamic environment influences the adoption of a model used for learning and decision making (Fig. 1c). Experiment 2 was similar to Experiment 1 except that reward probabilities assigned to different objects were not generalizable and could not be estimated accurately by combining the reward values of their features. 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. 1c).

Finally, in Experiments 3 and 4, we increased the dimensionality of the environment to examine the effect of a small and moderate dimensionality reduction resulting from feature-based learning. Reward probabilities, however, were fixed throughout both of these two experiments and reward values assigned to features were not fully generalizable to objects. This design allowed us to study the influence of dimensionality reduction and lack of generalizability on model adoption (Fig. 1c).

During Experiments 1 and 2, subjects completed 768 trials of a two-alternative choice task where on each trial, they selected between two colored shapes that provided reward probabilistically (Supplementary Figure 2a). These shapes were drawn from a set of four colored shapes each of which was assigned a specific reward probability. These reward probabilities, which we collectively refer to as the reward schedule, changed between blocks of 48 trials in order to generate environments with dynamic reward schedules. Overall, most subjects (68 out of 92 sessions) performed above the statistical chance level in both environments, indicating that they learned the values of options as they changed over time (Fig. 2a-c).

To examine the time course of learning, we computed the average probability of reward during each block of trials 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. 2b). Importantly, we did not find any significant change in achieving this performance over the course of the experiment (Supplementary Figure 3). This indicates that subjects did not use information from early reward schedules to predict future changes in reward schedules, which was challenging since the reward values for all four options changed between blocks.

To identify the learning model adopted 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 (Methods). We found that the coupled feature-based RL and feature-based RL with decay provided the best overall fits for the data in the generalizable environment (Table 1). Both of these feature-based RLs provided better fits than their corresponding object-based RLs, as measured by any of the goodness-of-fit indices (log likelihood, AIC, or BIC; Fig. 2c-e). By contrast, in the non-generalizable environment (Experiment 2), the coupled object-based RL and object-based RL with decay provided a significantly better fit than their corresponding feature-

based RLs (Fig. 2c-e). We found similar results when we considered each of four super-blocks separately (Supplementary Figure 4), indicating that the observed pattern of model adoption was not due to the use of different strategies early and late in the experiments.

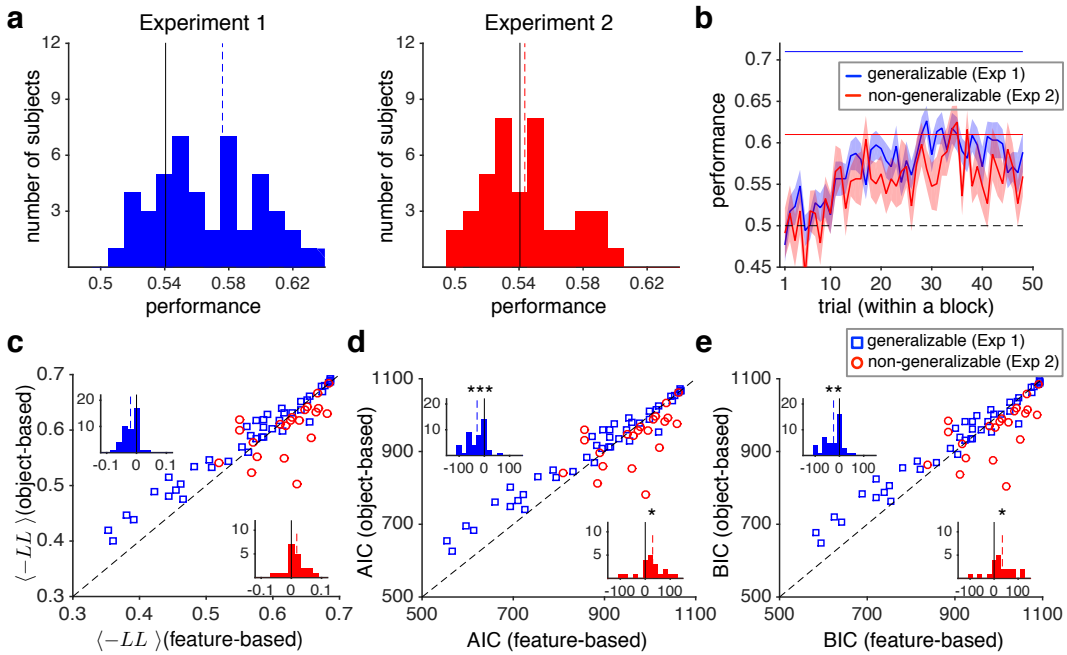


Figure 2. Dynamic reward schedules promote feature-based learning, whereas a lack of generalizability promotes object-based learning. (**a**) Performance or the average reward harvested by subjects during Experiments 1 (generalizable environment) and 2 (non-generalizable environment). Dashed lines show 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 in Experiments 1 and 2. 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 (-LL: negative log likelihood; AIC: Akaike information criterion; BIC: Bayesian information criterion). Plotted are the three measures of the goodness-of-fit 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.

Model	Coupled feature-based	Uncoupled feature-based	Feature-based with decay	Coupled object-based	Uncoupled object-based	Object-based with decay	
# pars.	5	5	6	4	4	5	
-LL	449.4±10.3	461.3±9.0	432.8±11.2	469.8±7.0	476.5±7.4	448.2±9.2	Exp. 1
AIC	908.9±20.6***	932.6±17.9***	877.6±22.5***	947.6±14.0	961.0 ±14.9	906.4±18.3	
BIC	932.1±20.6**	955.9±17.9**	905.5±22.5**	966.2±14.0	979.6±14.9	929.7±18.3	
-LL	492.0±6.3	496.0±5.1	476.7±7.8	487.6±5.1	499.4±4.3	462.8±7.7	Exp. 2
AIC	994.0±12.6	1002.1±10.1	965.4±15.6	983.3±10.3	1006.8±8.7	935.7±15.4*	
BIC	1017.2±12.6	1025.3±10.1	993.3±15.6	1001.8±10.3	1025.4±8.7	958.9±15.4*	
-LL	328.1±4.8	337.0±5.5	320.5±5.4	330.1±6.9	329.6±5.5	290.1±7.8	Exp. 3
AIC	666.2±9.6	684.0±10.9	653.0±10.8	668.2±13.9	667.1±11.0	590.2±15.7**	
BIC	687.9±9.6	705.6±10.9	679.0±10.8	685.5±13.9	684.4±11.0	611.9±15.7**	
-LL	377.7±6.5	376.0±6.8	331.7±8.9	409.3±4.1	410.6±4.0	349.2±6.9	Exp. 4
AIC	765.3±13.1***	762.0±13.6***	675.4±17.8***	826.6±8.3	829.2±7.9	708.4±13.9	
BIC	787.9±13.1***	784.5±13.6***	702.4±17.8**	844.6±8.3	847.2±7.9	730.9±13.9	

Table 1. Comparison of the goodness-of-fit measures in all experiments. Reported are the goodness-of-fit measures, negative log likelihood (-LL), Akaike information criterion (AIC), and Bayesian information criterion (BIC), averaged over all subjects (mean ± s.e.m.) for three feature-based RLs and their object-based counterparts for Experiments 1 to 4. The model providing the best fit in a given experiment and its object-based or feature-based counterpart are highlighted in cyan and orange, 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: $0.01 < p < 0.05$ (*), $0.001 < p < 0.01$ (**), and $p < 0.001$ (***).

Together, 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 accurate object-based learning.

Feature-based learning in static 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 Methods and Supplementary Figure 5). The purpose of the static environment was to isolate the influence of generalizability and dimensionality reduction on model adoption in the absence of changes in reward schedules studied in Experiments 1 and 2. Moreover, in order 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, resulting in non-generalizability of the environments (compare the average of values in individual columns or rows in Supplementary Figure 5a).

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. 3a). 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 over the course of the experiment from their reward probability estimates (see Methods). The fit of subjects' estimates revealed that the weight of the object-based approach, relative to the sum weights of the object-based and feature-based approaches, was much smaller than 0.5 during the first estimation block but gradually increased

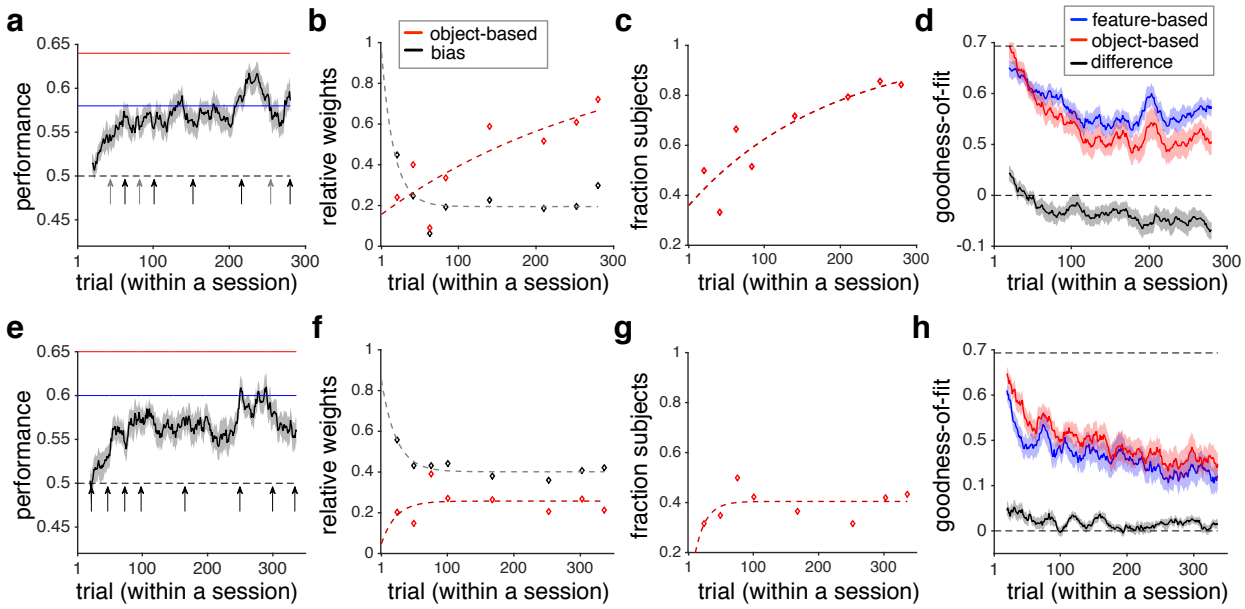


Figure 3. Transition from feature-based to object-based learning in static, non-generalizable environments. (**a**) The time course of performance during Experiment 3. The running average over time is computed using a moving box with the length of 20 trials. Shaded areas indicate s.e.m., and the dashed 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 the sum of the object-based and feature-based approaches, 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 who showed a stronger correlation between their reward estimates and actual reward probabilities than the probabilities estimated using the reward values of features. The dotted line shows the fit of data based on an exponential function. (**d**) Transition from feature-based to object-based learning revealed by the average goodness-of-fit over time. Plotted are the average negative log likelihood based on the best feature-based model, best 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.

over time (relative weight = 0.36, 95% CI [0.32 0.41] and 0.64, 95% CI [0.60 0.69] for the first two and last two estimates, respectively; Fig. 3b). In addition, the relative weight of bias (as an indication of subject's lack of discrimination between objects reward values) dropped to a small value early in the experiment. Similarly, correlation analysis revealed that during early estimation blocks, the estimates of only a small fraction of subjects were more correlated with actual reward probabilities than reward probabilities estimated using reward values of features, but this fraction increased over time (Fig. 3c). The results of these two analyses illustrated that subjects initially adopted feature-based learning and gradually switched to object-based learning.

We increased dimensionality of the environment in Experiment 4 in relation to Experiment 3 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, perhaps due to more frequent adoption of feature-based learning among our subjects (Fig. 3e). 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 (relative weight = 0.17, 95% CI [0.13 0.21] and 0.24, 95% CI [0.21 0.27] for the first two and last two estimates, respectively), while the relative weight of bias plateaued at a value larger than that in Experiment 3 (relative bias = 0.29, 95% CI [0.27 0.31] and 0.41, 95% CI [0.40 0.43] for the last two estimates in Experiments 3 and 4, respectively; Fig. 3f). Correlation analysis revealed a very similar pattern (Fig. 3g). All of these results suggest stronger feature-based learning compared to object-based learning when dimensionality or generalizability increased, because both these quantities increased in Experiment 4 relative to Experiment 3. As predicted by our framework (Fig. 1c), however, both higher generalizability and dimensionality should increase feature-based learning.

We also fit the data from Experiments 3 and 4 using various RL models in order to identify the model adopted by the subjects (see Methods). We found that object-based RL with decay provided the best overall fit in Experiment 3 (Table 1). Importantly, this model provided a better fit than its corresponding feature-based RL. Examination of the goodness-of-fit over time illustrated that the object-based learning model provided a better fit, particularly later in the experiment (Fig. 3d). The difference between the quality of the fit of the object-based and feature-based models in early (1-100) and late (100-280) trials

$(\langle -LL_{object-based} + LL_{feature-based}\rangle_{early} - \langle -LL_{object-based} + LL_{feature-based}\rangle_{late} =$ 0.019±0.050 (mean±std)) was significantly larger than zero (two-sided sign-test; $p < .05$). This indicates a transition from using feature-based to object-based learning. In Experiment 4, however, feature-based RL with decay provided the best overall fit (Table 1) and the fit of this model was better than the corresponding object-based learning model throughout the experiment (Fig. 3h). Overall, the results based on fitting the choice behavior were consistent with the results based on subjects' reward estimates.

Together, we found that during both Experiments 3 and 4, subjects transitioned from feature-based to object-based learning. However, an increase in the dimensionality accompanied by larger generalizability in Experiment 4 further biased the behavior toward feature-based learning.

Influence of attention on feature-based learning. Although the main goal of our study was to identify factors influencing how humans adopt feature-based versus object-based learning, our design also allowed us to examine how attention may influence learning. In all our experiments, the two features of options provided different amount of information. In principle, subjects could differentially attend to various features, resulting in unequal learning or in assigning different weights to the two features when constructing reward values during decision making. Therefore, we examined possible attentional effects by fitting choice behavior with a feature-based model with decay that has two separate learning rates for the more and less informative features. By design, this model can also assign different weights to the two features. We expected that attention would result in a larger learning rate and/or weight for the more informative relative to the less informative feature.

Comparison of the learning rates for the more and less informative features did not reveal any significant difference across all subjects (Fig. 4a-d). However, a significant number of subjects adopted a small learning rate (<0.001) for one of the two features during Experiments 3 and 4 (38% and 34%) but not during Experiments 1 and 2 (5% and 10%). Moreover, only during Experiments 3 and 4, those small learning rates were adopted more often for the less informative than the more informative feature (χ^2 (1) = 0.39, 0.72, 9.34, and 9.68 for equality of proportions during Experiments 1 to 4, respectively; $p < .01$ for Experiments 3 and 4 only). These results

indicate that during Experiments 3 and 4, subjects more frequently attended to and learned about the more informative compared to the less informative feature.

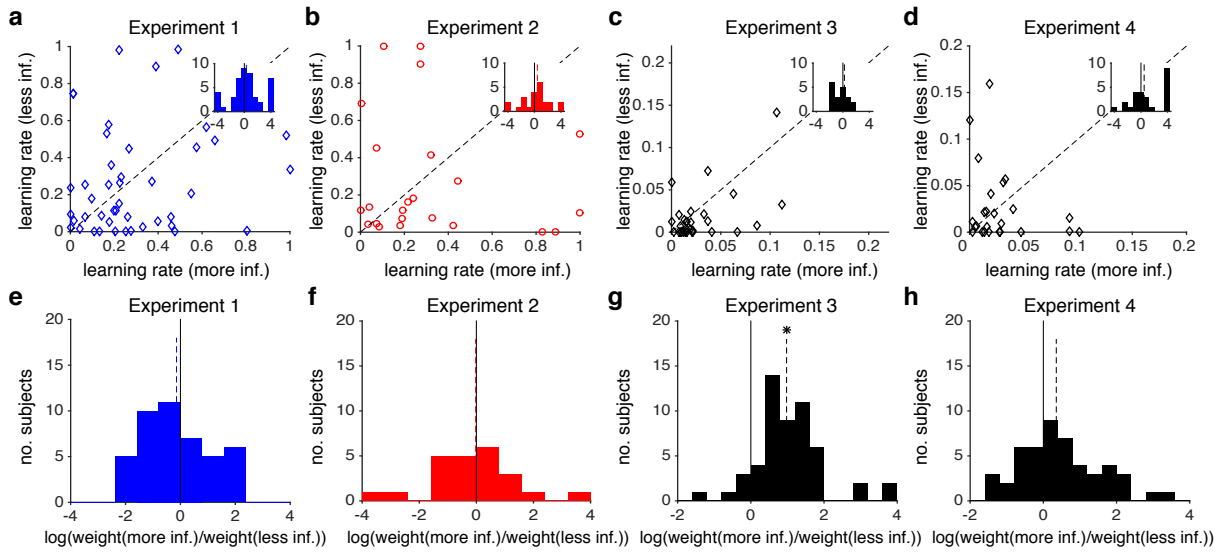


Figure 4. Dependence of the learning rates and assigned weights on the informative-ness of the features. More subjects adopted small learning rates for the less informative feature during Experiments 3 and 4, and more subjects assigned a larger 'attentional' weight to the more informative feature during Experiment 3. (**a-d**) The learning rate for the less informative feature (non-informative in the case of Experiment 2) is plotted versus that of the more informative feature for each subject in a given experiment. The insets show the histogram of the log ratio of the learning rates for the more informative to that of the less informative feature. The dashed lines show the medians and the solid gray lines indicate zero. There was no significant difference between the two learning rates in any of four experiments; however, more subjects adopted a small learning rate (<0.001) for the less informative feature during Experiments 3 and 4. (**e-h**) Plotted is the histogram of the log ratio of the weight assigned to the more informative feature to that of the less informative feature for each subject in a given experiment. The dashed lines show the medians and the solid grays line indicate zero. The star in panel **e** shows that the weights assigned to the two features were significantly different from each other. There was only evidence for a larger weight assigned to the more informative feature during Experiment 3.

We also examined the weights individual subjects assigned to the two features in order to detect any attentional bias in construction of objects' reward values from the reward values of their features. We did not find any evidence that subjects assigned a larger 'attentional' weight to the more informative feature in Experiments 1, 2, and 4 (Fig. 4e-h). During Experiment 3, however,

subjects assigned a larger weight to the more informative feature (two-tailed sign-rank test, p = .0003; Fig. 4g). Overall, these results indicate that in high-dimensional or static environments, subjects' choice behavior and learning were more strongly influenced by the information provided by the more informative feature, which could be due to deployment of attention on this feature.

To summarize our experimental results, we found that in dynamic environments, human subjects adopted feature-based learning even when this approach did not reduce dimensionality. Subjects switched to learning individual option values (object-based learning) when the combination of features' 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 constructed two network models in order to qualitatively capture our experimental observations and to gain insights into neural mechanisms for model adoption during learning in dynamic, multi-dimensional 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. 5a-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 values of presented objects or their features at any given time (see Methods for more details). 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 occurs 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. 5a). 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 and the final decision-making circuit. This increases the influence of the set of value-encoding neurons that was more likely responsible for making the final correct 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 final incorrect choice and the final decision-making circuit. This decreases the influence of the set of value-encoding neurons that was more likely responsible for making the final incorrect choice.

By contrast, the HDML model utilizes a signal-selection circuit to determine which set of the value-encoding neurons contains a stronger signal, and updates connections from the OVE and FVE neurons to their corresponding signal-selection accordingly. In this model, signal strength is defined as the difference between the reward values of the two options based on the output of OVE or FVE neurons. The model uses only the output of the set with a stronger signal to make the final decision on a given trial (Fig. 5b). Subsequently, only the strength of connections between the set of value-encoding neurons 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 Methods for more details).

We used the two network models to simulate learning during our behavioral experiments. We first examined the behavior during Experiment 1 using a generalizable, 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. 5c). This occurred 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.

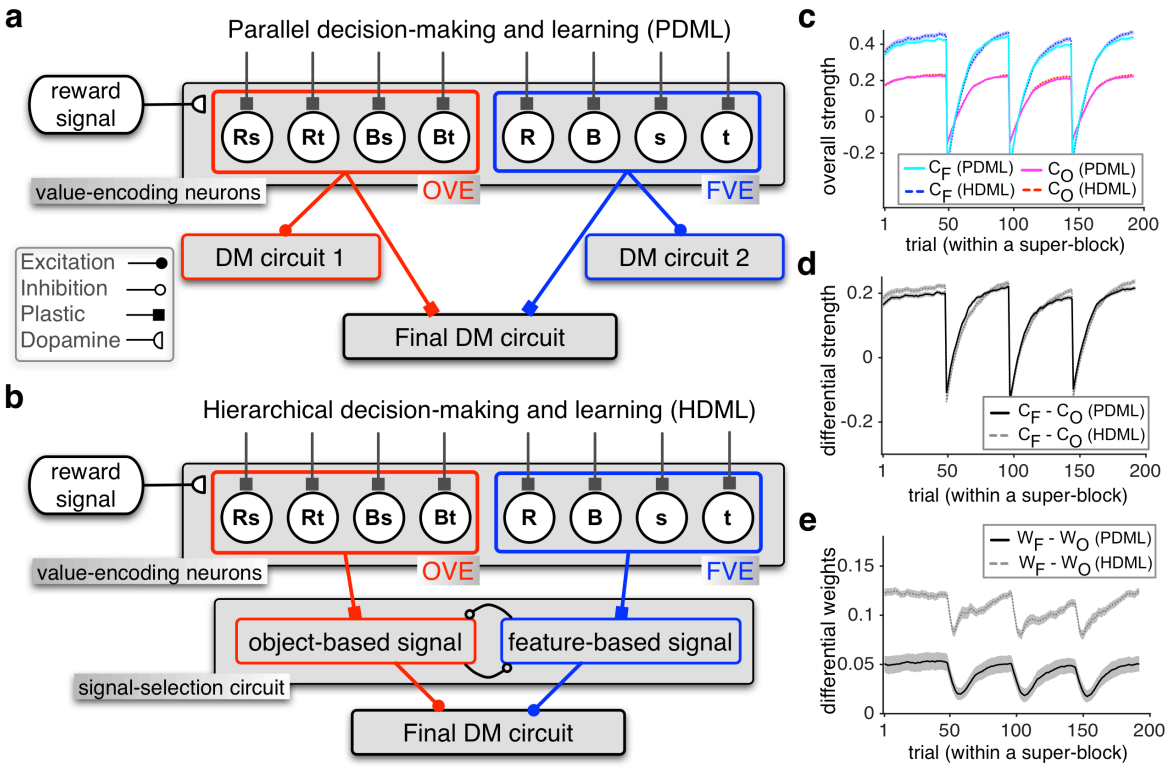


Figure 5. Architectures and performances of two alternative network models for multi-dimensional, decision-making tasks. (**a-b**). Architectures of the PDML (a) and the HDML (b) models. (**c**) The time course of the overall strengths of plastic synapses between OVE and FVE neurons and the final DM 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 difference in the overall weights of the two sets of value-encoding neurons on the final decision ($W_F - W_O$) for the same set of simulations shown in panels **c** and **d**.

These faster updates enabled the FVE neurons to signal a correct response more often than the OVE neurons following each change in reward probabilities (Fig. 5d). We also computed the overall weight of the feature-based and object-based approaches on the final choice, W_F and W_O, respectively (see Methods). The difference between these two weights, (W_F - W_O), was positive in both models even though it decreased after each reversal, indicating that both models assigned a larger weight to feature-based than to object-based reward values. However, this effect was greater in the HDML compared to the PDML model (Fig. 5e).

To study how generalizability and frequency of changes in reward probabilities (volatility) affect model adoption, we used the two network models to simulate various environments with different levels of generalizability and volatility. These environments were constructed by varying the relationship between the reward value of each object and the reward values of its features, and changing the block length, i.e. the number of trials where reward probabilities were fixed (see 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 slightly higher in all environments (Fig. 6a, d, g). More importantly, the difference in the strength of connection from FVE and OVE neurons to the next stage of processing ($C_F - C_O$) was more strongly modulated by generalizability and volatility in the HDML compared to the PDML model. This indicated that HDML was better able to adjust the strength of connections from value-encoding neurons (Fig. 6b, 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 rather than object-based reward values (larger $W_F - W_O$) as the environment became more generalizable or volatile (Fig. 6c, f, i). Overall, these results demonstrated that, although both models were able to perform the task, the HDML model exhibited higher performance and 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.

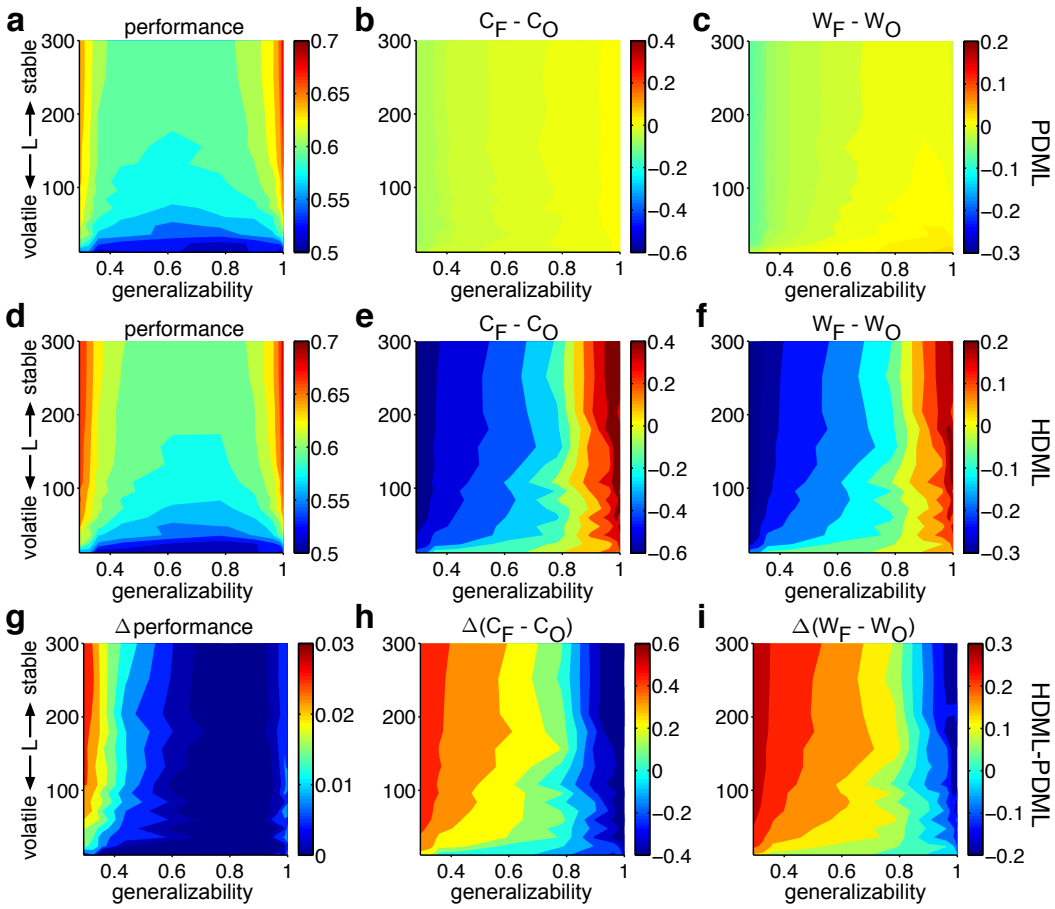


Figure 6. The effects of generalizability and volatility (i.e. frequency of changes in reward probabilities) on the models' behavior. (**a**) Performance of the PDML model in various environments with different levels of volatility and generalizability. The color map shows the performance (average harvested 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. (**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, (C_F -C_O), and (W_F– W_O) in the HDML and PDML models.

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 dimensionality is a discrete number, we simulated choice behavior in two different environments with different number of feature instances (three or four) resulting in dimensionality $D = 3^2$ and $D = 4^2$. We also changed the level of generalizability

across different environments (see Methods). Consistent with simulation results for Experiments 1 and 2 presented in Figure 6, an increase in generalizability caused both models to assign higher weights to feature-based rather than object-based reward values, but this effect was much stronger for the HDML model (larger positive slopes in Figure 7e-f compared with Figure 7b-c). An increase in dimensionality further biased both models to assign more weight to feature-based compared to object-based reward values.

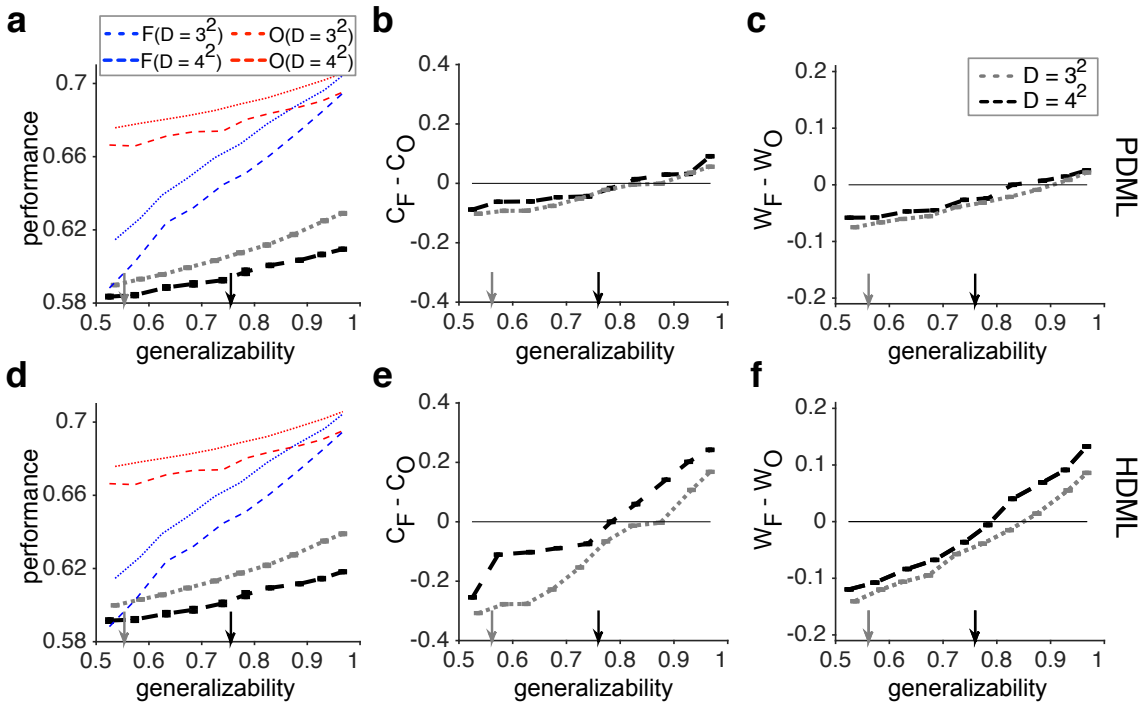


Figure 7. The effects of dimensionality and generalizability on the models' behavior. Simulations of the PDML and HDML models are shown in **a-c** and **d-f**, respectively. (**a, d**) Performance of the models as a function of the generalizability index for two environments with 9 (gray) and 16 (black) objects, respectively. The dotted red and blue curves show the maximal performance for object-based (O) and feature-based (F) models, respectively, for $D = 3^2$, as in Experiment 3. The dashed curves show the results for $D = 4^2$, as in Experiment 4. 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).

Overall, the simulation results for the two alternative network models illustrated that the HDML model exhibited higher performance and stronger adjustment to task parameters and reward statistics in the environment. These results indicate that that hierarchical decision-making and learning might provide a more plausible mechanism for adopting the model for learning in dynamic, multi-dimensional environments.

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 confirmed these predictions and demonstrated that dynamic environments tend to favor feature-based learning, because this learning 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. These results suggest that the main driver for adopting heuristic feature-based learning is increasing adaptability without compromising precision; that is, to overcome the adaptability-precision tradeoff (APT).

The APT sets an important constraint on learning reward values in a dynamic environment where they change over time. One solution to mitigate the APT is to adjust learning over time via metaplasticity (Farashahi et al., 2017; Khorsand and Soltani, 2017). Nevertheless, even with adjustable learning, the APT still persists and becomes more critical in multi-dimensional environments, since the learner may never receive reward feedback on many unchosen options and feedback on chosen options is quite limited. Importantly, adopting heuristic 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, Chang, Park, Heinzle, & Haynes, 2012).

Our results could explain why learning in young children, which is limited by the 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 well in simple tasks such as the dimension-switching task (Zelazo, Frye, & Rapus, 1996). Interestingly, this inability has been attributed to a failure to inhibit attention to the previously relevant or rewarding feature (Kirkham, Cruess, & Diamond, 2003). Here, we propose an alternative possibility that by focusing on a single feature such as color, children could evaluate reward outcomes of chosen options based on color and thus increase their learning speed. Moreover, by choosing a favorite color, they can further reduce dimensionality by decreasing the number of feature instances/categories to two; favorite and non-favorite color. Thus, our results explain that choosing a favorite color not only reduces dimensionality but also increases adaptability without compromising precision.

Even though rules used for the heuristic feature-based approach are only partially generalizable in the real world, this lack of generalizability may not prevent humans from adopting feature-based learning for a few reasons. First, simply due to chance, the level of generalizability is larger for a higher dimensionality if there is at least one informative feature in the environment. Second, reward values of features can be learned separately for different domains (e.g. color of fruits and color of cars), since each domain individually is more generalizable. Finally, non-generalizability may never be detected due to a very large number of features and options in real world. Together, we suggest that feature-based learning could provide a "fast and frugal way" for learning in the real world (Gigerenzer & Goldstein, 1996).

Heuristic feature-based learning is computationally less expensive and more feasible than object-based learning, since it can be achieved using a small number of value-encoding neurons with *pure feature selectivity*, namely neurons that represent the reward value in a single dimension, such as color or shape. By comparison, object-based learning requires myriad mixed selectivity neurons tuned to specific combinations of various features. Thus, 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. 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, 2015; Logothetis, Pauls, & Poggio, 1995).

Our computational and experimental results also provide a few novel neural predictions. First, they predict that learning about reward in dynamic environments could depend more strongly on value-encoding neurons with pure feature selectivity, since activity or representation of such neurons can be adjusted more frequently over time due to more updates per feedback. Second, 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 (Khorsand, Moore, & Soltani, 2015; Soltani, Khorsand, Guo, Farashahi, & Liu, 2016). Third, they 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 and become obsolete. Finally, the complexity of reward value representation should be 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 becomes more feasible and thus, more complex representation of reward values will emerge. These novel predictions could be tested in future experiments.

As our computational modeling suggests that learning based on feature-based and object-based approaches occurs simultaneously in two separate circuits, and arbitration between the two forms of learning might be required (Lee, Seo, & Jung, 2012; Lee, Shimojo, & O'Doherty, 2014; Seo, Donahue, Cai, & Lee, 2014). Our modeling results show that such arbitration could happen via competition between two circuits based on the strength of signals in each circuit. Although we could not directly fit experimental data using these two models due to the over-fitting problem and large between-subject variability, 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 our human subjects. In the HDML model, the best sources of information were identified to make decisions, and weights associated with the selected sources were successively updated according to reward feedback. The hierarchical structure allows the HDML model to reduce noise in decision making by ignoring the less informative value-coding network on each trial. Together,

we find that 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 also been suggested as an underlying mechanism behind multi-attribute decision making (Hunt, Dolan, & Behrens, 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 our model 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.

Despite the fact that naturalistic learning from reward feedback entails options with overlapping features, 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 et al., 2014; Niv et al., 2015; Vaidya, 2015; Wilson & Niv, 2012; Wunderlich et al., 2011). A few of these studies have found that learning in a multi-dimensional environment relies on constructing a simplified representation of the stimuli via attending to one feature and ignoring others (Eldar et al., 2013; Niv et al., 2015). This finding, however, could be biased by the experimental paradigm in which only one of the three features was informative and objects did not predict reward based on the combination of their features. By contrast, reward probabilities in our experiments were assigned to specific options and no single feature alone could accurately predict reward on all options. Consequently, we found some evidence for attentional bias on one of the two features whereas no feature was ignored. More specifically, although the rate of learning did not always increase for the more informative features, we found that more subjects adopted a small learning rate for the less informative feature during Experiments 3 and 4. The absence of similar effects in Experiments 1 and 2 could be due to a higher dimensionality or static reward schedules in Experiments 3 and 4. Moreover, only during Experiment 3, subjects also assigned more weight on the more informative feature during decision making, consistent with allocating more attention to that feature. Attending to only a subset of "relevant" features is both inevitable and crucial for learning and decision making in high-dimensional environments (Payne, Bettman, & Johnson, 1993; Tversky, 1972). However, in order to identify the relevant features in dynamic environments, values of multiple features should be updated in parallel over time.

In conclusion, we show that a tradeoff between adaptability and precision could explain why humans adopt feature-based learning, especially in dynamic environments. Moreover, our results suggest that neurons with pure selectivity could be crucial for learning in dynamic environments and could provide a missing framework for understanding how heterogeneity in reward representation emerges (Wallis & Kennerley, 2010; Donahue & Lee, 2015).

References

Barraclough, D.J., Conroy, M.L. & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature neuroscience*, *7*(4), 404-410.

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

Botvinick, M.M. (2012). Hierarchical reinforcement learning and decision making. *Current opinion in neurobiology*, *22*(6), pp.956-962.

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.

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.

Donahue, C. H., & Lee, D. (2015). Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. *Nature Neuroscience, 18,* 295-301.

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

Farashahi, S., Donahue, C. H., Khorsand, P., Seo, H., Lee, D., & Soltani, A. (2017). Metaplasticity as a Neural Substrate for Adaptive Learning and Choice under Uncertainty. *Neuron, 94(2),* 401-414.

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.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gross, C. G., Rocha-Miranda, C. E. , & 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. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning: data mining, inference and prediction. *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.

Ito, M. & Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience*, *29*(31), 9861-74.

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. *Frontiers in Psychology, 6,* 155.

Khorsand, P., & Soltani, A. (2017). Optimal structure of metaplasticity for adaptive learning. *BioRxiv,* 129619; https://doi.org/10.1101/129619.

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.

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.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Ribas-Fernandes, J.J., Solway, A., Diuk, C., McGuire, J.T., Barto, A.G., Niv, Y. & Botvinick, M.M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*(2), 370-379.

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.

Seo, H., Cai, X., Donahue, C. H., & Lee, D. (2014) Neural correlates of strategic reasoning during competitive games. *Science*, 346(6207) 340-3.

Soltani, A., Khorsand, P., Guo, C., Farashahi, S., & Liu, J. (2016). Neural substrates of cognitive biases during probabilistic inference. *Nature Communications*, *7*, 11393.

Soltani, A., & Koch, C. (2010). Visual saliency computations: mechanisms, constraints, and the effect of feedback. *The Journal of Neuroscience*, *30*(38), 12831–12843.

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. *Current Opinion in Neurobiology*, *18*, 209–216.

Soltani, A., & Wang, X.-J. (2010). Synaptic computation underlying probabilistic inference. *Nature Neuroscience*, *13*(1), 112–119.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*(4), 281–299.

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

Wallis J. D., & Kennerley S. W. (2010). Heterogeneous reward signals in prefrontal cortex. *Current Opinion in Neurobiology*, 20(2):191-8.

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.

Methods

Framework for adoption of feature-based versus object-based learning. To test our hypothesis that feature-based learning is adopted mainly to overcome the adaptability-precision tradeoff, we first developed a general framework for understanding model adoption during learning in dynamic, multi-dimensional environments. The decision maker's task is to learn the reward values of a set of options via reward feedback after selecting one of two alternative options on each trial. We simulated the behavior of two contrasting learners in this task: the object-based and feature-based learner. The object-based learner directly estimates the value of individual objects via reward feedback. By contrast, the feature-based learner estimates the reward values of all feature instances (e.g. red, blue, square, or triangle) by updating the reward values associated with all the features of the object for which reward feedback is given. This learner then combines the reward values of feature instances to estimate the overall reward value of individual objects.

Assuming that options/objects have m features, each of which can have n different instances, there are n^m possible objects in the environment. For example, if an object has two features (m = 2), color and shape, and there are three colors and three shapes (n = 3), there would be nine (3^2) possible objects in the environment. We first constructed an environment by assigning a probability of reward to each object based on the feature instances of that object. More specifically, n feature instances were assigned with a set of equally-spaced (in log scale) likelihood ratios (LR) in all m dimensions. The minimum and maximum values of LRs 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 instance j ($j = 1,\ldots,n$) of feature i ($i = 1,\ldots,m$). The LR for a given object a, $LR(O_a)$, was determined by multiplying the LRs of all features of that object: $LR(O_a) = \prod_{i=1, for\ F_{ij}\ present\ in\ O_a}^{m} LR(F_{ij})$. Finally, the probability of reward on each object was then computed by transforming the object's LR to the probability of reward: $p_r(O_a) = LR\ (O_a)/(1 + LR\ (O_a))$. These reward probabilities are referred to as the *fully-generalizable reward matrix*.

Although each object is assigned a reward value, the feature-based learner could instead use the reward value for each feature instance (e.g. red, green, triangles, squares) to estimate reward

values of objects in two steps. First, the reward value for a given feature instance (e.g. red) can be computed by averaging the reward values of all objects that contain that feature instance (e.g. all red objects): $\overline{p_r}(F_{ij}) = (1/n^{m-1}) \sum_{O_a\ contains\ F_{ij}} p_r(O_a)$. Second, an estimated reward value, $\tilde{p}_r(O_a)$, can be generated by combining the reward values of features using the Bayes theorem:

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

These estimated reward probabilities constitute the *estimated reward matrix* based on features. The order of probabilities in the estimated reward matrix is similar to that of the fully-generalizable reward matrix, whereas the exact values may differ slightly (diamonds in Supplementary Figure 1).

By randomly shuffling the elements of the fully-generalizable reward matrix in all feature dimensions except one, which we call the informative feature, we generated environments with different levels of generalizability. We used the correlation between the 'shuffled' reward matrix and estimated reward matrix (i.e. correlation between the actual reward value of options and the estimated reward value of options based on their features) to define a generalizability index. Based on this definition, the generalizability index 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 order of the estimated reward values of options based on their features is identical to the order of actual objects' values. With less generalizability, the order is not the same (reflected in a smaller generalizability index) and the difference between the estimated reward values based on features and actual objects' values increases (Supplementary Figure 1).

The task for the decision maker is to learn the reward value of options/objects via reward feedback in order to choose between two alternative options on each trial. The object-based learner directly estimates the reward value of all objects using 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 a, $r(t)$ 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 feature-based learner estimates the reward value of individual feature instances (e.g. red, green, triangles, squares), $V_{F_{ij}}(t)$, using the same update rule as in Equation 2, but applying to all features of the chosen object. This learner then combines the reward values of feature instances to compute reward values of each option (Eq. 1). Therefore, the object-based learner only updates one value function after each feedback whereas the feature-based learner updates the reward value of all feature instances of the selected object.

To measure how well a learner that uses the object-based approach 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:

$$S_O(t) = \frac{1}{n^m \times (n^m - 1)} \sum_{a=1}^{n^m} \sum_{b=1}^{n^m} \left(V_{O_a}(t) - V_{O_b}(t)\right) sign(p_r(O_a) - p_r(O_b)) \quad \text{(Eq. 3)}$$

where $p_r(O_a)$ is the probability of reward on object a. The differential signal for the feature-based learning model, $S_F(t)$, was computed by replacing $V_{O_a}(t)$ in the above equation with the estimated reward value $\tilde{V}_{O_a}(t)$, which was computed by replacing $\overline{p_r}(F_{ij})$ in Equation 1 with $V_{F_{ij}}(t)$. Therefore, the differential signal measures how reward values estimated by a given model correctly differentiate between actual reward values of objects.

By comparing the time courses of the differential signal for the object-based and feature-based learners (using the same learning rate and similar initial conditions), we computed the time at which the object-based learner carries a stronger differential signal than the feature-based learner (the 'cross-over point'). A larger cross-over point indicates the superiority (better performance) of the feature-based relative to the object-based learning for a longer amount of time, whereas a zero cross-over point indicates that the object-based learning is always superior.

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 and/or 2 (33 subjects performed in both experiments). This resulted in the behavioral data from N = 51 and 41 subjects 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 0.5406 (equal to 0.5 plus 2 times s.e.m., based on the average of 608 trials after excluding the first 10 trials of each block in Experiment 1 or 2). This resulted in the exclusion of data from 8 of 51 sets in Experiment 1, and 20 of 41 sets in Experiment 2. The data from the remaining 64 sessions was used for further analysis (N = 43 and 21 for Experiments 1 and 2, respectively). For Experiment 3, 36 additional subjects were recruited (20 females) and a performance threshold of 0.5447 (equal to 0.5 plus 2 times s.e.m., based on the average of 500 trials after excluding the first 30 trials of each session) was used to exclude subjects whose performance was indistinguishable from chance (N = 9). 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 0.5404 (equal to 0.5 plus 2 times s.e.m., based on the average of 612 trials after excluding the first 30 trials of each session) was used to exclude subjects whose performance was indistinguishable from chance (N = 11). 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 Department of Psychological and Brain Sciences at Dartmouth College. 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 each of these experiments, subjects completed two sessions (each session composed of 384 trials and lasting about half an hour) of a choice task during which they selected between a pair of objects on each trial (Supplementary Figure 2a). 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.

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 consisting of some permutations of four reward probabilities [0.1, 0.3, 0.7. 0.9], were used. In eight of these schedules, a generalizable rule could be used to predict reward probabilities for all objects based on the combinations of their feature values (Supplementary Figure 2b). In the other eight schedules, no generalizable rule could be used to predict reward probabilities for all objects based on the combinations of their feature values (Supplementary Figure 2c). For example, the schedule notated as 'Rs' indicates that red objects are much more rewarding than blue objects, square objects are more rewarding than triangle objects, and color (uppercase 'R') is more informative than shape (lower case 's'). 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 reward values of individual objects and their features (e.g. red square was the most rewarding object, but red triangle was less rewarding than blue triangle). In other words, the non-generalizable reward schedules were 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 their environments were composed of reward schedules with generalizable and non-generalizable rules, respectively (Supplementary Figure 2d, f). In both experiments, as the subjects moved between blocks of trials, reward probabilities for the informative features were reversed without any changes in the average reward probabilities for the less informative and non-informative feature in Experiments 1 and 2, respectively. For example, going from Rs to Bs changes the more informative feature instance from red to blue. The changes in reward probabilities occurred without any cue to the subject and created dynamic environments. In addition, the average reward probabilities for the less informative or non-informative feature changed (e.g., from Bs and Rs to Bt and Rt) every four blocks (super-blocks; Supplementary Figure 2e, g). Each subject performed the experiment in each environment once, where either color or shape was consistently more informative. The more informative feature was randomly assigned and counter-balanced across subjects to

minimize the effects of intrinsic color or shape biases. The order of experiments was randomized for subjects who preformed both Experiments 1 and 2.

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 (Supplementary Figure 5a; one of nine possible objects with a 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 from a total of 8 patterns and 8 shapes. 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 (Supplementary Figure 5c). All subjects completed five blocks of estimation trials throughout the task (after trials 42, 84, 140, 210, and 280 of the choice task), and 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.

Selection of a given object was rewarded (independently of the other presented object) based on a reward schedule with a moderate level of generalizability such that reward probability of some individual objects could not be determined by combining the reward values of their features. 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 was not. Although the informative feature (e.g. pattern and shape in Supplementary Figure 5a right and left panels, respectively) was on average predictive of reward, this prediction was not generalizable. That is, some objects that

contained the most rewarding feature instances were still less rewarding than objects that did not contain these feature instances. For example, S1P3 in the left panel of Supplementary Figure 5a was less rewarding than S1P2. Finally, the average reward probability of the objects with the same non-informative feature instances (e.g. S1P1, S1P2, S1P3 in Supplementary Figure 5a left panel) was always 0.5. This reward schedule ensured that subjects would not be able to predict reward probability accurately for all objects based on the combination of their feature values. Similar 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 instances 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 (Supplementary Figure 5b; 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 values for features were similar for Experiments 3 and 4.

Data analysis. We utilized the information subjects provided during estimation trials of Experiments 3 and 4 in order to examine how they determined the reward values of objects, using two alternative methods. First, we used linear regression to fit the estimates of reward probabilities as a function of the following regressors: actual reward probabilities assigned to each object (object-based term); the reward probabilities estimated based on the combination of the reward values of features (feature-based term) using the Bayes theorem (Eq. 1); and a constant. The constant (bias) in this regression model quantifies subjects' overall bias in reporting probabilities. Moreover, the relative weight of the bias term to the sum of all regressors 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 the reward values of features (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 based on the following equation:

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

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

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

$$\text{logit}\, P_{O1}(t) = \big(V_{O1}(t) - V_{O2}(t)\big)/\sigma + bias \qquad \text{(Eq. 6)}$$

where P_{O1} is the probability of choosing object 1, V_{O1} and V_{O2} are the reward values of the two presented objects, *bias* measures a response bias toward the left or right option to capture the subject's location bias, and σ is a parameter measuring the level of stochasticity in the decision process.

Feature-based RL models. In this group of models, the reward value (probability) of each object is computed by combining the reward values of the features of that object, which are 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 Experiments 3 and 4, the two alternative objects were always different in both features. In Experiments 1 and 2, however, the two alternative objects could have a common feature instance (e.g. both are blue) and updating the reward value of this common feature could be problematic. Indeed, we found that the fit of choice behavior based on a feature-based model which always updates the reward values of both features of the selected object on each trial was worse than that of all other tested models (data not shown). Therefore, in the feature-based models presented here, only the reward value of the unique feature is updated when the two alternative options have a common feature on a given trial.

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

$$\text{logit}\, P_{O1}(t) = w_{shape}\big(V_{shapeO1}(t) - V_{shapeO2}(t)\big) + w_{color}(V_{colorO1}(t) - V_{colorO2}(t)) + bias$$

(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, *bias* measures a response bias toward the left or right option to capture the subject's location bias, and w_{shape} and w_{color} determine the influence of the two features on the final choice. Note that these weights can be assumed to be learned over time through reward feedback (as in our models; see below) or could reflect differential processing of the two features due to attention.

RL models with decay. Additionally, we investigated the effect of 'forgetting' the reward values of unchosen objects or feature(s) by introducing decay of value functions (in the uncoupled models) which has been shown to capture some aspects of learning (Barraclough et al., 2004; Ito & Doya, 2009), especially in multi-dimensional tasks (Niv et al., 2015). More specifically, the reward values of unchosen objects or feature(s) decay 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.

Estimating attentional effects. Attention could influence how reward values of two features determine choice and how they are updated over time. Therefore, in order to distinguish these two roles of attention, we estimated learning rates as well as the 'attentional' weights separately for the less and more informative features. By design, the feature-based models assign two different weights to the two features before combining them to make a choice (Eq. 7). We also extended the feature-based model with decay to include separate learning rates for the less and more informative features. For fitting of choice behavior in Experiments 3 and 4, we adopted two sets of weights for the first and second session of the experiments since two different sets of stimuli were used in these two sessions.

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. 5a-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 values 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 based on reward feedback. More specifically, 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. 5a). If the final choice was rewarded, the model increases the strength of connections between the set or sets of value-encoding neurons 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. 5b). 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.

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 an object or feature (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 instances 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:

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

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

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. Note that similarly to the models used for fitting, only the reward value of the unique feature of the selected object was updated when the two alternative objects had a common feature.

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:

$$C_m(t+1) = C_m(t) + q_+\big(1 - C_m(t)\big),\ \ if\ r(t) = 1, and\ pool\ m\ choice = final\ choice$$

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

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

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 et al., 2006; Soltani & Wang, 2006, 2008, 2010). This difference is proportional to the difference in the overall fraction of 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):

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

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:

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

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:

$$\text{logit}\, P(O_1) = (F_{O1} - F_{O2})/\sigma\,,\ \ if\ OVE\ signal\ is\ selected$$

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

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

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

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

where m is the selected signal.

Models 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. 6). 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 (Fig. 7), we changed the levels of generalizability by randomly shuffling some of the elements of the fully-generalizable

reward matrices with two values of dimensionality (3^2 and 4^2). The reward probabilities were fixed over the course of these simulations, as in the real Experiments 3 and 4.

Models parameters. Both models have six parameters: 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 overall behavior of the models did not qualitatively depend on these parameters.

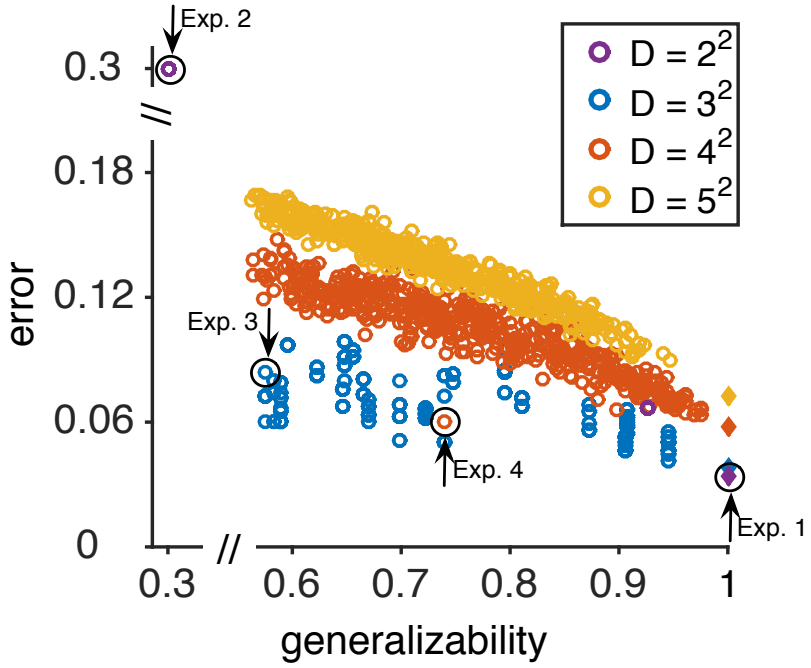
Assessment of models' response to different environments. We assessed how the two models responded to properties of the environment, in terms of generalizability, volatility, and dimensionality, in three different ways. First, we measured performance, defined as the average harvested reward in a given environment. Second, we measured 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. Finally, we measured the difference in the overall weights that object-based and feature-based reward values exert on the final choice in each model.

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. In both models, however, the overall influence of the object-based or feature-based values on choice also depends on how signals encoded in plastic synapses onto the OVE and FVE neurons can differentiate between objects reward values. We computed such a 'differential signal' (S) for the object-based reward values 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

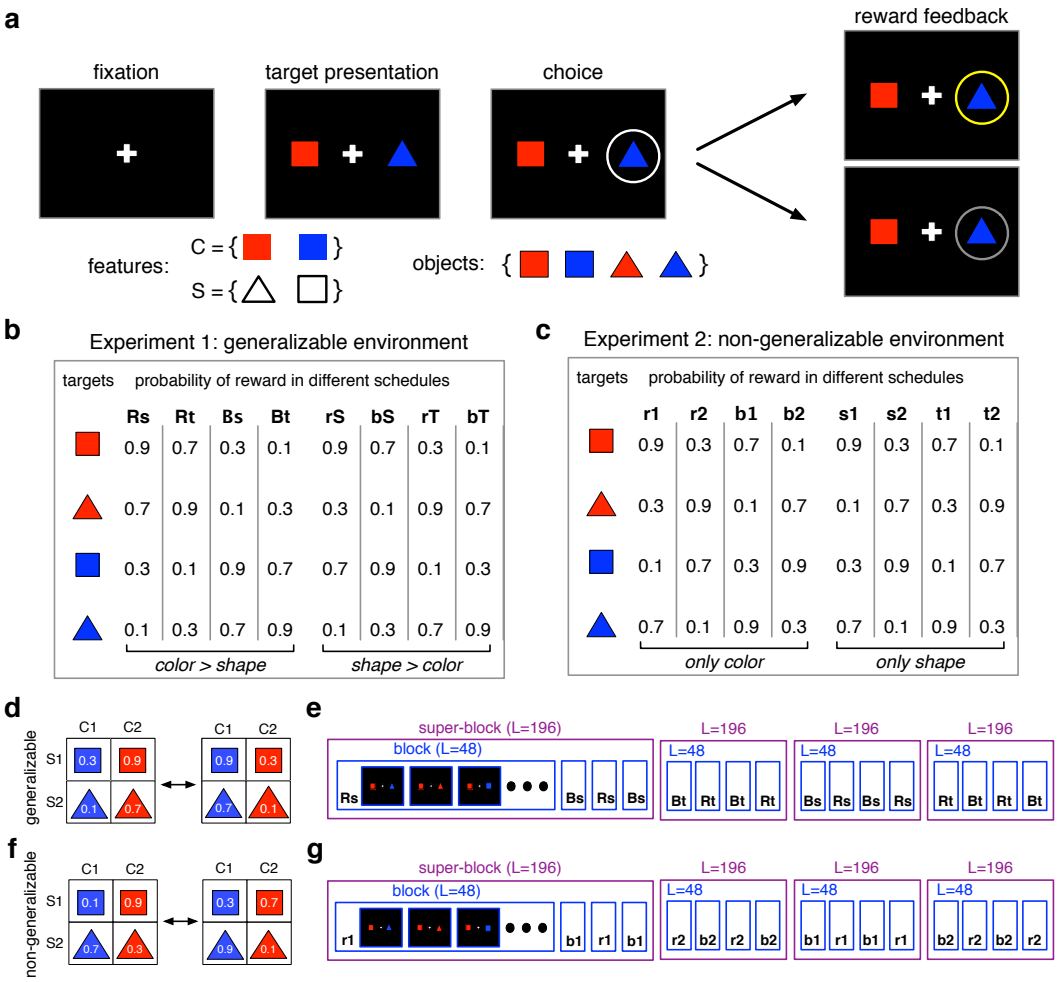
reward values for objects based on the synaptic strengths for synapses onto FVE neurons selective to shape and color ($F_{shape,i}(t)$ and $F_{color,i}(t)$) and Equation 1.

Finally, the overall weight of the object-based and feature-based values on the final choice was computed by the product of the differential 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. More specifically, 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 $C_F(t) \times S_F(t)$.

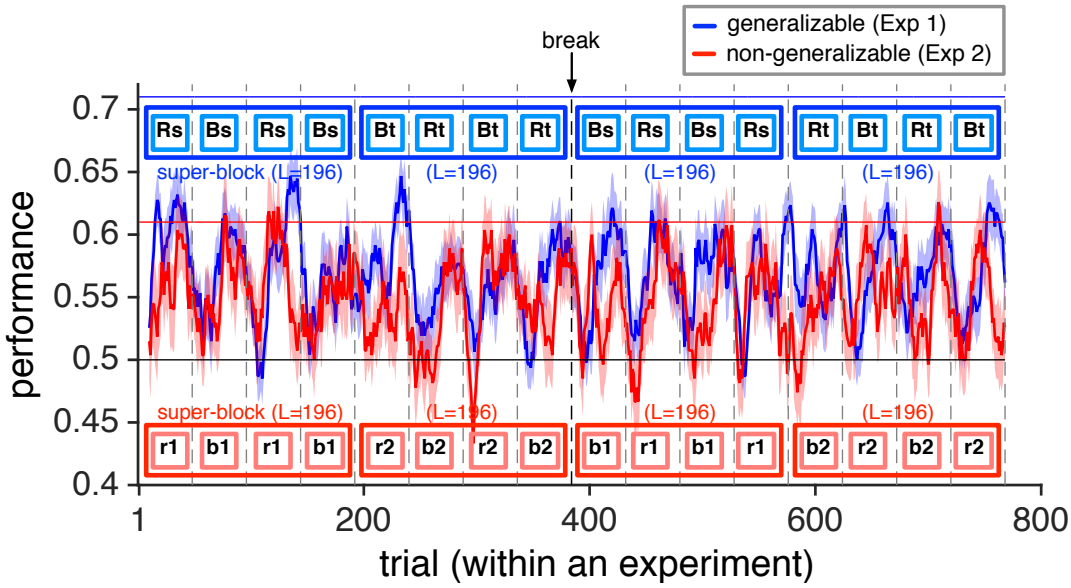
Supplementary Figures



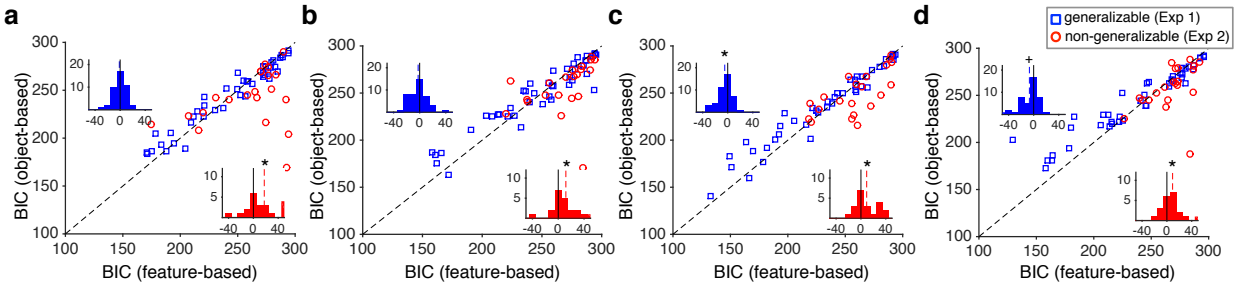
Supplementary Figure 1. Error in the estimation of reward probabilities using the reward values of features. Plotted is the mean of absolute difference between the estimated reward probabilities based on features (Eq. 1) and the actual reward probabilities, as a function of the generalizability index separately for environments with different values of dimensionality. The error increases with smaller generalizability and with larger dimensionality. Error values for fully generalizable environments are plotted with filled diamonds. The black circles indicate error values for Experiments 1 to 4. The generalizability and error for reward matrices used in Experiments 3 and 4 are different from environments with similar dimensionality because of the removal of a few non-informative objects in these experiments.



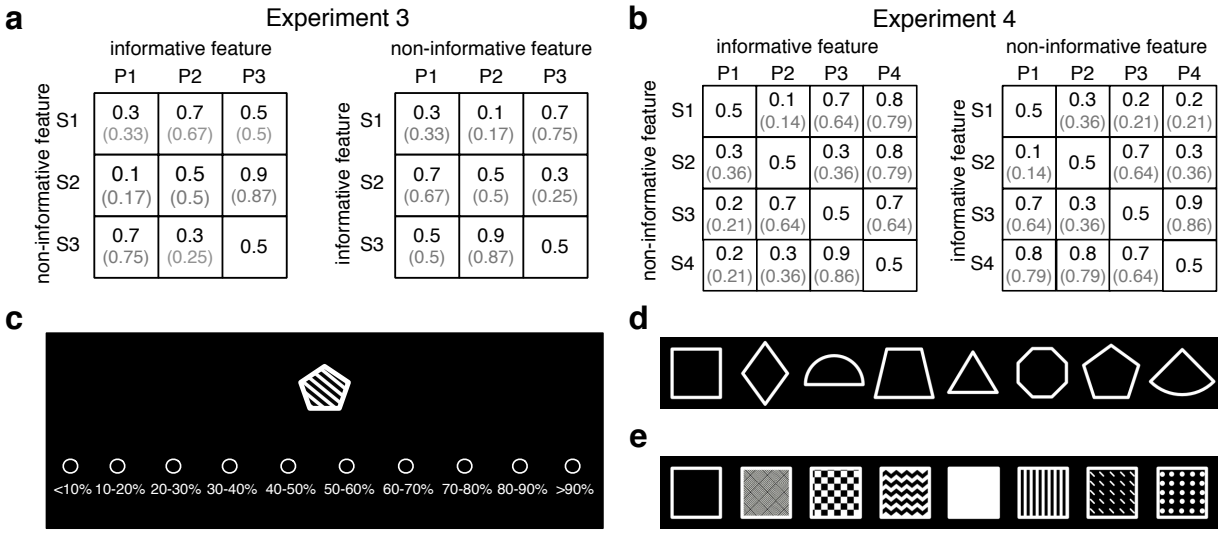
Supplementary Figure 2. Timelines and reward schedules of Experiments 1 and 2. (**a**) On each trial, the subject chose between two objects (colored shapes) and was provided with reward feedback (reward or no reward) on the chosen object. The inset shows the set of all objects used during Experiments 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 instances 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.



Supplementary Figure 3. Time course of learning during the entire course of Experiments 1 and 2. Plotted is the average harvested reward on a given trial within an experiment across all subjects (the shaded areas indicate s.e.m.). The black solid line shows chance performance, and the blue and red solid lines show the maximum performance in the generalizable and non-generalizable environments, respectively. The blue and red boxes show example sequences of reward schedules and super-blocks in the two environments. Notations for reward schedules are the same as in Supplementary Figure 1. Overall, performance increased over the course of each block and dropped after reversal in both experiments. Importantly, there was no evidence for using different strategies in early and late blocks of trials.



Supplementary Figure 4. Comparison of the goodness-of-fit based on the best object-based and feature-based model in each super-block of Experiments 1 and 2. Panels **a** to **d** show the results for super-blocks 1 to 4, respectively. Plotted are the BIC 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 (plus sign) shows that the median is significantly different from zero at $p < 0.05$ using a two-tailed (one-tailed), sign-rank test. These results show that, from early on in the experiments, subjects were more likely to adopt a feature-based approach in the generalizable environment and an object-based approach in the non-generalizable environment, and that our results were not driven by two types of behavior during early and late parts of the experiments.



Supplementary Figure 5. Reward probabilities and objects used in Experiments 3 and 4. (**a**) 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 the reward probabilities assigned to all objects could not be determined by combining the reward values of their features (non-generalizable). Numbers in parentheses show the actual probability values used in the experiment due to limited resolution for reward assignment. For the set on the left, the pattern was on average more informative about reward, whereas shape alone was not informative. The opposite was true about the right set. 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. (**b**). 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. (**c**) 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. (**d**) 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). (**e**) 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.