Contributions of statistical learning to learning from reward feedback

Aryan Yazdanpanah[1]*, Michael Chong Wang[1]*, Ethan Trepka[1,2], Marissa Benz[1], Alireza Soltani[1]

Affiliation

[1]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA

[2]Department of Neuroscience, Stanford University, Stanford, CA, USA

**These authors contributed equally to this work*

Corresponding author: Alireza Soltani (alireza.soltani@dartmouth.edu)

Classification: Social Sciences: Psychological and Cognitive Sciences; Biological Sciences: Psychological and Cognitive Sciences

Keywords: learning, decision making, reward, adaptive behavior, naturalistic learning

Manuscript information: 5 figures, 5 tables, 4 supplementary figures, 45 pages
Abstract: 150 words
Total word count, excluding the Abstract, Methods, References, and figure legends: 5500 words

Competing interests: The authors declare no competing interests.

Authors contributions

AS conceived the idea for the experiment. AY, MCW, ET, and AS designed the experiment. MCW and ET developed the computer code for running the experiment. AY, MCW, and MB conducted the experiment. AY, MCW, and AS developed the computational models. AY and MCW implemented the computational models and performed model simulations. AY, MCW, and AS analyzed the experimental data and simulations. AY, MCW, and AS wrote the first draft of the manuscript. All authors edited the manuscript.

Acknowledgements

We would like to thank Daniela Garcia and Elizabeth Frey for helping with the design of the experiment, and Elizabeth Frey for developing the code for running the experiment. This work was supported by a National Science Foundation CAREER Award (BCS1943767) to A.S.

Abstract

Natural environments are abundant with patterns and regularities. These regularities can be captured through statistical learning, which strongly influences perception, memory, and other cognitive functions. By combining a sequence-prediction task with an orthogonal multidimensional reward learning task, we tested whether detecting environmental regularities can also enhance reward learning. We found that participants used regularities about features from the sequence-prediction task to bias their behavior in the learning task. Fitting choice behavior with computational models revealed that this effect was more consistent with attentional modulations of learning, rather than decision making. Specifically, the learning rates for the feature with regularity were higher, particularly when learning from forgone options during unrewarded trials, demonstrating that statistical learning can intensify confirmation bias in reward learning. Overall, our findings suggest that by enhancing learning about certain features, detecting regularities in the environment can reduce dimensionality and thus mitigate the curse of dimensionality in reward learning.

Significance statement

Natural environments are filled with detectable patterns and regularities, which, once identified through statistical learning, engage our attentional system and significantly influence multiple cognitive functions. This study explores whether these processes can enhance reward learning in high-dimensional environments with limited reward feedback. Using a novel experimental paradigm and computational methods, we discovered that detecting regularities in specific stimulus features increases learning rates for those features, especially for unrewarded, forgone options. Our findings suggest that identifying environmental regularities can improve feature-based learning and mitigate the curse of dimensionality.

Introduction

While constantly changing, natural environments exhibit numerous spatial and temporal regularities across multiple timescales. For instance, trees blossom in the spring, and their fruits change color from green to red during the summer. Detecting and learning such regularities is ecologically crucial, as it enhances predictability of the environment and reduces its complexity, allowing animals to navigate and adjust their behavior more efficiently (Gibson, 1969). Not surprisingly, our brain is adept at identifying relationships between events, cues, and stimuli, a process termed statistical learning (Saffran et al., 1996; Turk-Browne et al., 2009). This ability has been shown to enhance other cognitive functions such as object categorization (Turk-Browne et al., 2010), short-term memory (Umemoto et al., 2010), auditory perception (Saffran et al., 1999), language acquisition, and social cognition (Saffran & Kirkham, 2018).

Ultimately, survival also hinges on the ability to learn and select options and actions that yields reward, commonly referred to as reward learning. In naturalistic environments, however, reward learning can be very challenging because numerous choice options and/or actions precede a single reward feedback while each option possesses multiple features or attributes. This complexity causes the required sample size for effective learning to grow exponentially, leading to the so-called curse of dimensionality (Sutton & Barto, 2018). One possible solution to mitigate the curse of dimensionality is to learn the predictive values of features to estimate reward values associated with individual objects or choice options (Farashahi et al., 2017; Leong et al., 2017; Niv et al., 2015). Such feature-based learning provides a good candidate for learning reward contingencies in multidimensional reward environments as it can reduce dimensionality without compromising too much precision (Farashahi et al., 2017, 2018, 2020). But how are these features are identified in the first place, especially when choice options possess many features?

Unlike reward learning that requires specific events (e.g., reward feedback), statistical learning can happen continuously as we receive the stream of sensory stimuli from our environment. Indeed, humans can automatically extract multiple statistical features of regularities without explicit instruction (Fiser & Aslin, 2002). Furthermore, it has been suggested that statistical learning may be considered a form of associative learning (Turk-Browne et al., 2009). These raise the question of whether statistical learning can improve reward learning by mitigating the curse of dimensionality.

A possible mechanism for this could be the impact of statistical learning on attention: by focusing attention on certain features to enhance learning from those features while suppressing others, statistical learning effectively leads to dimensionality reduction. Despite this, statistical relationships among cues, objects and events, and their reward associations could be unrelated in naturalistic environments, which may cause statistical learning to bias reward learning towards features that are not particularly informative about reward outcomes. Nonetheless, if regularities provide information about environmental changes that predict the availability of food, shelter, etc., their detection and utilization can ultimately enhance survival. For example, statistical learning enables the detection of color changes in fruits across seasons, typically shifting from green to red. This detection could focus attention on colors for feature-based, reward learning from shades of green to red. Although color is not universally predictive of taste, nutrients, or calories across all fruits (i.e., it is not generalizable), this feature-based learning can significantly speed up the learning process and assist in identifying other fruits.

Therefore, we hypothesized that detecting regularities and corresponding statistical learning about certain features of choice options can enhance feature-based learning from reward feedback on those features. To test this hypothesis, we designed a novel experimental paradigm involving two orthogonal tasks: a sequence-prediction task to engage statistical learning, and a multidimensional reward learning task to study more naturalistic reward learning. During the sequence-prediction task, human participants were asked to predict the next stimulus in a sequence by identifying regularities within a feature. This was done to create temporal regularities in one feature (the manipulated feature) to focus attention on that feature, while keeping the sequence of the other feature random. In the multidimensional reward learning task, participants learned the predictive value of a different set of stimuli by receiving reward feedback after choosing between pairs of stimuli. Importantly, each stimulus was associated with a reward probability that could be partially estimated using its features, with both features were equally informative in predicting the reward outcome. We employed multiple analyses to examine how regularity manipulation influenced learning and choice behavior. Additionally, to identify underlying mechanisms, we developed various computational models and used these models to fit choice data on a trial-by-trial basis.

Our results show that temporal regularities in the sequence-prediction task influenced reward learning for a different set of stimuli that shared similar features. Specifically, the feature with temporal regularity was learned more strongly (larger learning rates), even though both features provided equal information about reward outcomes. We found little evidence that regularity manipulation influenced decision making. Overall, these findings support our hypothesis that statistical learning shapes reward learning, with significant implications for learning in naturalistic environments.

Results

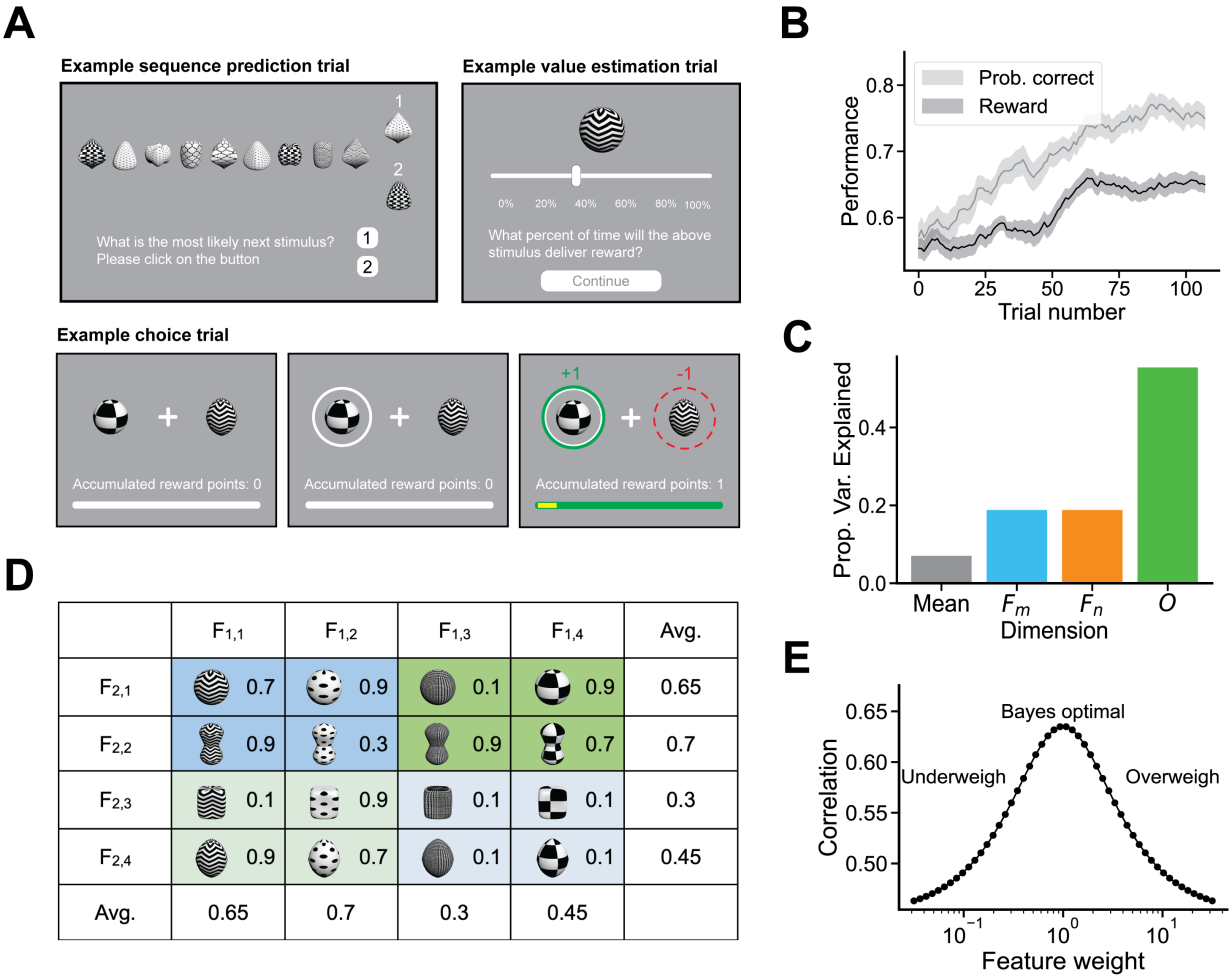
To investigate whether environmental regularities can influence reward learning to mitigate the curse of dimensionality, we designed a novel experimental paradigm that involved performing two independent tasks: a multidimensional reward-learning task interleaved with a sequence-prediction task (**Fig. 1A**). During the multidimensional learning task, human participants learned about the predictive value of 16 visual stimuli, each with two features (shape and pattern), after selecting between a pair of stimuli and receiving reward feedback on each trial. In addition to these choice trials, the learning task required participants to provide their estimates about the reward probabilities associated with each stimulus (estimation trials) similar to our previous studies (Farashahi et al., 2017; Farashahi & Soltani, 2021).

The interleaved sequence-prediction task was specifically designed to create temporal regularities related to a certain feature of the stimuli. To introduce such regularities, during each trial of the sequence-prediction task, participants were presented with a sequence of visual stimuli and were required to identify the next stimulus in the sequence. Crucially, while these stimuli had visual features similar to those in the learning task, they were distinct from the set of stimuli used in that task. Moreover, the sequence of stimuli included repetitions (doublets, triplets, or quadruplets) of one of the two features, referred to as the manipulated feature, making the next stimulus predictable based on that feature. In contrast, the other feature was presented in a random fashion and thus had no predictive power about the next stimulus. To become acquainted with the task, each participant initially performed 10-20 trials of the sequence-prediction task. During the actual experiment, a bout of sequence-prediction trials was presented before every 8 choice trials, and a bout of estimation trials followed every 32 choice trials, for a

total of 4 bouts of estimation trials (see **Methods** for more details). Sequence-prediction trials were interleaved with choice trials throughout the experiment to ensure the full impact of our manipulation was maintained.

Crucially, despite the difference in temporal regularity imposed during the sequence-prediction task, the reward schedule of the learning task was symmetric with respect to feature values. This made both features equally informative about the reward outcome (**Fig. 1D**). However, the reward schedule was not entirely generalizable, as both features and object identities independently accounted for parts of the total variance in reward probabilities associated with the 16 stimuli used in the learning task. (**Fig. 1C**). As a result, the estimation of reward probabilities based on feature value results in a certain amount of errors (Farashahi et al., 2017). Nonetheless, the feature-based learner should combine the weights of both features equally as over-weighing and under-weighing each feature will lead to further error in estimation and more suboptimal behavior (**Fig. 1E**).

Despite the inherent difficulty of the reward learning task, which involved 16 stimuli, and the interleaved sequence-prediction task, the majority of participants successfully completed the experiment. Specifically, 57 out of the 78 total participants achieved a performance level above chance in both the learning and sequence-prediction tasks, and/or their estimates of reward probability associated with each stimulus were significantly correlated with the assigned reward probabilities (see **Methods** for more details on exclusion criteria). The time course of performance of the participants indicated that learning continued throughout the experiment and only plateaued after 80 (choice) trials (**Fig. 1B**).



	$F_{1,1}$	$F_{1,2}$	$F_{1,3}$	$F_{1,4}$	Avg.
$F_{2,1}$	0.7	0.9	0.1	0.9	0.65
$F_{2,2}$	0.9	0.3	0.9	0.7	0.7
$F_{2,3}$	0.1	0.9	0.1	0.1	0.3
$F_{2,4}$	0.9	0.7	0.1	0.1	0.45
Avg.	0.65	0.7	0.3	0.45	

Figure 1. Experimental paradigm, overall performance, and the structure of the reward environment. (**A**) The experiment consisted of three types of trials. In sequence-prediction trials, participants were presented with a sequence of stimuli and asked to predict the next stimulus in the sequence. In choice trials, participants chose between a pair of stimuli and received reward feedback, which included the potential reward they would have received if they had selected the unchosen stimulus. In value-estimation trials, participants provided their estimates of the reward probabilities associated with all 16 stimuli. (**B**) Time course of learning based on choice trials. The black and gray curves represent the average percentage of reward per trial and the proportion of trials in which the more rewarding stimulus was selected across all participants. The curves were smoothed using a moving average with a window of 20 trials. The shaded areas show the s.e.m. (**C**) Variance decomposition of the reward schedule. Both features explained the same amount of variance in the reward outcome, but a substantial amount of variance was only explained by the combination of features, allowing us to disentangle feature-based and object-based learning. (**D**) Reward schedule. The table displays the reward probabilities associated with an example set of 16 choice options (stimuli) defined by the conjunction of two features ($F_{i,j}$ corresponds to instance j of feature i). Reward probabilities are symmetric for the two features. To facilitate learning, stimuli in the dark blue cells were consistently paired with those in the light blue cells, while stimuli in the dark green cells were paired with those in the light green cells. The average values represent the mean reward probabilities for stimuli that share a common feature. (**E**) Correlation between the true reward probabilities and the estimated reward probabilities using a weighted combination of feature values. The optimal combination requires equal weight on both features. Over-weighing or under-weighing one feature leads to suboptimal estimations.

Effects on regularity manipulation on choice behavior and value estimations

To test whether temporal regularities in the environment can modulate reward-based learning via its effects on feature-based attention, we first analyzed participants' choice behavior and value estimations to search for signatures of attentional modulation. Our previous study, which used a similar multidimensional learning task but without the temporal regularity manipulation, suggests that the choice behavior of individuals who adopt feature-based learning can be predicted by a weighted combination of feature values (Wang & Soltani, 2023). Therefore, we fit a generalized linear mixed-effects model using the log ratio of the true feature and object values of the two options to predict participants' choices (**Table 1**).

Predictor	*b*	*SE*	*z*	*p*
Intercept	-0.01	0.03	-0.20	0.838
Block	0.01	0.03	0.37	0.708
F_m value	1.21	0.14	8.78	<0.001
F_n value	0.78	0.11	7.32	<0.001
O value	0.12	0.04	3.08	0.002
Block * F_m value	0.28	0.10	2.89	0.004
Block * F_n value	0.22	0.09	2.38	0.017
Block * O value	0.11	0.03	3.73	<0.001

Table 1. Results of fitting choice data with a generalized linear mixed-effects model based on log odds of feature and object values and block as independent fixed factors. F_m, F_n and O denote the manipulated feature, non-manipulated feature, and object, respectively. The block factor denotes whether the trial was in the first or second half of the session. Interaction between block and the value predictors reflect changes in knowledge about feature and object values due to learning.

We found that participants' choices were significantly influenced by the log of value ratio of the manipulated feature (b=1.20, SE=0.14, z=8.78, p<0.001), the log of value ratio of the non-manipulated feature (b=0.78, SE=0.11, z=7.32, p<0.001), and to a much lesser extent, the log of value ratio associated with the stimulus itself (b=0.12, SE=0.04, z=3.09, p=0.002). This indicates that participants utilized a mixture of feature-based and object-based learning.

Importantly, the contrast between the coefficients for the manipulated and non-manipulated features was significant ($\chi^2(1)$=7.34, p=0.007), suggesting that the values of the manipulated

feature had a greater effect on choice behavior. The best-fit logistic curves, which predict the probability of choosing an option based on the log odds of feature values between the options, also demonstrate this relationship. More specifically, the slope of these curves reflects the main effects in the generalized linear mixed-effects model, indicating that the manipulated feature exerted a stronger influence on choice behavior (**Fig. 2A**).

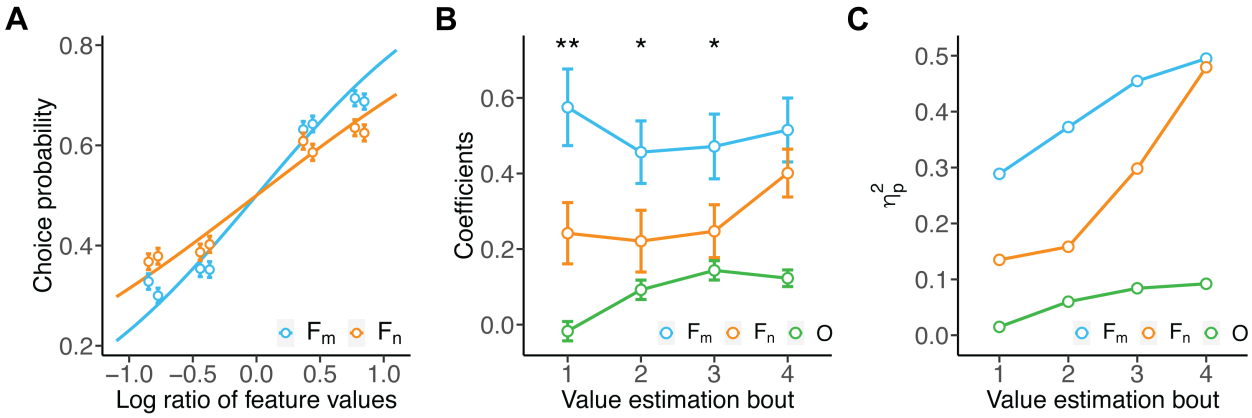


Figure 2. Choice and value estimation strategies during the learning task. (**A**) Results of mixed-effects logistic regression using the log ratio of feature values to predict choice. Each point shows the choice probability associated with a pair of stimuli, error bars show 1 s.e.m., and the lines are the best fit logistic curves. The slope for the manipulated feature (F_m) was significantly higher than that of the non-manipulated feature (F_n). (**B**) Results of a mixed-effects linear regression using feature values and object identity (O) to predict participants' estimated values across four bouts of estimation trials. Each point represents the fixed effects regression coefficient, with error bars indicating the standard error. Asterisks above the points indicate the significance of the coefficient contrast between the manipulated feature and non-manipulated feature (*p<0.05; **p<0.01). (**C**) Proportion of variance explained by different fixed effects in the mixed-effects linear regression used to predict participants' estimates based on feature values and object identity. Each point represents η_p^2 of each feature and the object. The manipulated feature accounted for more variance in the data compared to the non-manipulated feature, suggesting that it had a more significant effect on participants' estimation of values. This difference is reduced over time while the effect of object identity increases.

To gain a more detailed understanding of how participants computed and represented the reward values of different stimuli (i.e., the reward probability associated with each stimulus), we examined their value estimates for all stimuli during estimation bouts that occurred at four time points throughout the experiment. For this analysis, we employed linear mixed-effects models that used the log odds of features and objects values associated with individual stimuli to predict participants' estimations of the reward probability associated with each stimulus. We found that

for three out of four estimation bouts, the coefficients on the manipulated feature were significantly higher than the coefficients of the non-manipulated feature (**Table 2**).

Bout	*Predictor*	*b*	*SE*	*t*	*df*	*p*
1	Intercept	0.05	0.04	1.14	797.60	0.256
	F_m	0.58	0.10	5.68	72.74	<0.001
	F_n	0.24	0.08	2.98	84.45	0.004
	O	-0.02	0.03	-0.67	797.60	0.502
	F_m-F_n	0.33	0.12	2.76	108.99	0.007
2	Intercept	0.09	0.04	2.22	797.76	0.027
	F_m	0.46	0.08	5.52	83.20	<0.001
	F_n	0.22	0.08	2.70	84.01	0.008
	O	0.09	0.03	3.59	797.76	<0.001
	F_m-F_n	0.24	0.11	2.21	116.88	0.029
3	Intercept	0.12	0.04	3.15	797.36	0.002
	F_m	0.47	0.09	5.50	80.68	<0.001
	F_n	0.25	0.07	3.53	98.19	<0.001
	O	0.14	0.03	5.61	797.36	<0.001
	F_m-F_n	0.22	0.10	2.24	110.19	0.027
4	Intercept	0.12	0.03	3.45	797.92	<0.001
	F_m	0.52	0.08	6.10	73.93	<0.001
	F_n	0.40	0.06	6.34	92.80	<0.001
	O	0.12	0.02	5.58	797.92	<0.001
	F_m-F_n	0.11	0.10	1.17	104.97	0.245

Table 2. Results of fitting participants' estimated values with a linear mixed-effects model based on log odds of feature and object values as independent fixed factors, separately for each bout of estimation trials. F_m, F_n, and O denote the manipulated feature, non-manipulated feature, and object, respectively. F_m-F_n denotes the result of the contrast test.

In our previous study involving multidimensional learning tasks, we found that participants were unlikely to choose different stimuli that contain a certain feature with equal probability (Wang & Soltani, 2023). This led to value estimates that systematically differ from the predictions solely based on feature values, which are defined as the average value of stimuli sharing a feature Therefore, without assuming these feature values, we also performed the analysis of variance on a linear mixed-effects model with the feature identities and their interactions as the independent

variables (**Table 3**). We found that for all trials, η_p^2 for the manipulated feature was higher than that of the non-manipulated feature. Overall, this suggests that although both features had identical predictive values, participants placed greater weight on the value of the manipulated feature when estimating the values of stimuli. Moreover, in addition to rapidly acquiring the feature values, participants gradually learned the more accurate stimulus values, as reflected in the increased influence of object identity on their estimations (**Table 3**). This supports the previously reported transition from feature-based to object-based learning (Farashahi et al., 2017, 2020).

Bout	*Predictor*	*F*	*Num. df*	*Den. df*	*p*	η_p^2
1	F_m	14.42	3	106.56	<0.001	0.29
	F_n	6.11	3	117.59	<0.001	0.13
	O	0.82	9	483.47	0.602	0.01
2	F_m	19.28	3	97.51	<0.001	0.37
	F_n	8.27	3	131.84	<0.001	0.16
	O	3.46	9	486.80	<0.001	0.06
3	F_m	27.59	3	99.24	<0.001	0.45
	F_n	18.70	3	132.06	<0.001	0.30
	O	5.14	9	504.81	<0.001	0.08
4	F_m	32.18	3	98.54	<0.001	0.49
	F_n	39.46	3	128.58	<0.001	0.48
	O	5.80	9	514.66	<0.001	0.09

Table 3. Results of fitting participants' estimated values with a linear mixed-effects model based on feature and object identities as fixed factors, separately for each bout of estimation trials. F_m, F_n and O denote the manipulated feature, non-manipulated feature, and object, respectively. The object effect is the same as interaction between the features because each stimulus can be identified as a conjunction of two features.

Overall, these results suggest that participants made choices and estimated values based on a weighted combination of feature values, while giving more weight to the manipulated feature. To understand the emergence of this bias toward the manipulated feature, we next explored how choice on a given trial depended on choice and reward outcomes on the preceding trials.

Effects of regularity manipulation on credit assignment and response to reward feedback

Previous studies on reward learning have demonstrated that both previous choice and reward outcomes can influence choice behavior (Barraclough et al., 2004; Kim et al., 2009; Witkowski et al., 2022). A general finding is that the reward outcome strongly influences future choices such that options that are rewarded are more likely to be selected again, whereas the absence of a reward typically leads to a switch from the previously chosen option. This pattern is commonly known as the 'win-stay-lose-switch' (WSLS) strategy and can be used to measure adjustments in learning and choice behavior (Trepka et al., 2021). Moreover, the selection of an option can also increase the likelihood of choosing that same option in the future independently of reward outcome, a phenomenon often referred to as 'choice autocorrelation' (CA). Both WSLS and CA can be captured using logistic regression to predict choice on a given trial based on the choice and reward outcome on the preceding trial (Katahira, 2018; Lau & Glimcher, 2005; Wang & Soltani, 2023).

Here, we expanded this approach in multiple ways to more thoroughly examine learning from multidimensional choice options. First, we defined WSLS and CA based on the individual features of the choice options (visual stimuli) to measure reward credit assignment and choice strategies according to these features. Second, to model counterfactual learning, we incorporated both the reward and choice histories associated with the forgone options in the regression model. This was possible because reward feedback was provided for both chosen and unchosen stimuli in our study. Finally, to capture transitions from feature-based learning to mixed feature-based and object-based learning (Farashahi et al., 2017, 2020), we fit separate models to the trials in the first and second halves of the session. Using the outcome mixed-effects logistic regression model, we analyzed choice behavior of participants during the learning task (**Table 4**).

First block

Predictor	*b*	*SE*	*z*	*p*
Intercept	-0.01	0.04	-0.32	0.749
F_m chosen WSLS	0.21	0.06	3.37	<0.001
F_m chosen CA	0.19	0.09	2.18	0.029
F_m forgone WSLS	0.30	0.06	4.87	<0.001
F_m forgone CA	-0.24	0.07	-3.38	<0.001
F_n chosen WSLS	0.18	0.06	3.01	0.003
F_n chosen CA	0.01	0.08	0.11	0.914
F_n forgone WSLS	0.05	0.06	0.86	0.392
F_n forgone CA	-0.18	0.08	-2.41	0.016
O chosen WSLS	0.08	0.15	0.56	0.578
O chosen CA	0.18	0.15	1.20	0.23
O forgone WSLS	-0.04	0.18	-0.23	0.82
O forgone CA	-0.04	0.16	-0.23	0.816

Second block

Predictor	*b*	*SE*	*z*	*p*
Intercept	-0.00	0.04	-0.00	0.999
F_m chosen WSLS	-0.05	0.07	-0.79	0.431
F_m chosen CA	0.61	0.11	5.62	<0.001
F_m forgone WSLS	0.11	0.06	1.70	0.088
F_m forgone CA	-0.33	0.08	-4.03	<0.001
F_n chosen WSLS	0.03	0.07	0.40	0.688
F_n chosen CA	0.37	0.08	4.51	<0.001
F_n forgone WSLS	0.15	0.06	2.25	0.024
F_n forgone CA	-0.01	0.08	-0.19	0.853
O chosen WSLS	0.53	0.17	3.14	0.002
O chosen CA	-0.28	0.17	-1.65	0.099
O forgone WSLS	0.38	0.16	2.38	0.018
O forgone CA	-0.42	0.16	-2.61	0.009

Table 4. Result from generalized linear mixed-effects model predicting choice on a trial-by-trial basis based on reward and choice outcomes on the previous trial. The model incorporates individual features and stimulus identity, separately for the first and second block of the experiment. WSLS: win-stay-lose-switch, CA: choice autocorrelation. Chosen and forgone refer to chosen and unchosen option or feature on the previous trial.

This analysis provided several new insights into learning from multidimensional stimuli. Firstly, we observed a positive CA for the chosen option and a negative CA for the forgone option. A positive CA for the chosen option indicates that participants were inclined to select the previously chosen option again, or an option that shared one of its features. Conversely, a negative CA for the forgone option implies that participants tended to avoid selecting the previously unchosen option, or an option that shared a feature with it. Secondly, we found a transition of learning strategies from feature-based to object-based learning. In the first half of the trials, participants' choices were only significantly influenced by reward and choice histories associated with the features of previous options, but not the combination of them (i.e., the object). In the second half of the study, the WSLS strategy based on the features of choice options was minimal, whereas CA based on option features was high. This indicates that by the second half of the experiment, participants may have developed strong value representations based on the features of the stimuli. At the same time, the reward and choice histories associated with the object (stimulus identity) significantly influenced their choice behavior. These findings suggest a shift from learning based on individual features to learning based on the stimulus/object (combination of features). Importantly, this shift was not because object values were learned more slowly than feature values. Rather, in the first half of the trials, rewards were either not associated with the object at all, or they had minimal impact on subsequent choices.

Lastly, participants' choices were more influenced by the reward and choice history of the manipulated feature. In the first half of the experiment, the regression coefficients for the reward and choice histories of the manipulated feature were all significantly different from zero. In addition, the magnitudes of these coefficients were all larger than those for the non-manipulated feature (**Fig. 3**). This effect was particularly evident in learning about the forgone option, where participants showed significant WSLS behavior based on the manipulated feature (b=0.30, SE=0.06, z=4.87, p<0.001), but not based on the non-manipulated feature. In the second half of the trials, participants' choices were more consistent based on the manipulated feature. This was indicated by a large positive CA with the chosen option and a large negative CA with the forgone options.

Overall, these results provide compelling evidence that detecting temporal regularity and subsequent statistical learning has a lasting effect on learning through feature-based attention.

They also confirm the dynamic shift from feature-based learning to object-based learning that has been observed in previous studies (Farashahi et al., 2017, 2020).

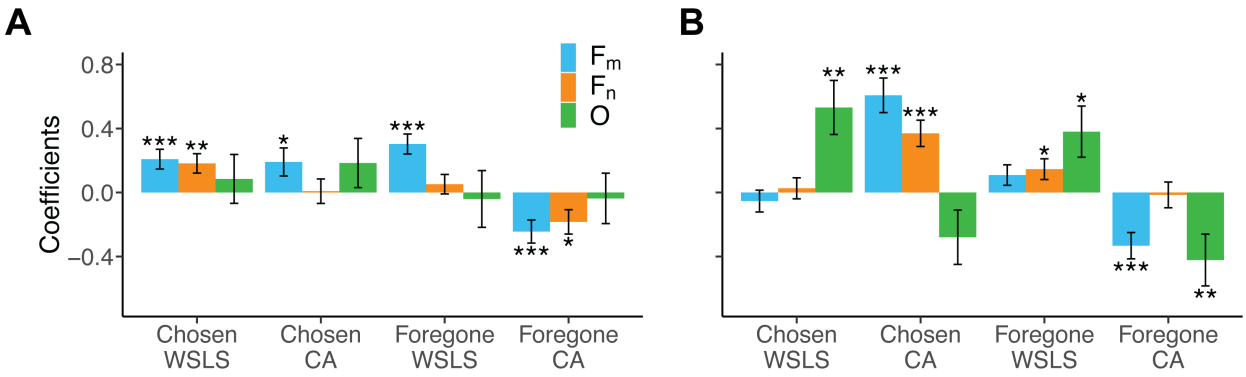


Figure 3. The influence of choice and reward outcomes on choice behavior estimated using a mixed-effects logistic regression model. WSLS: win-stay-lose-switch, CA: choice autocorrelation. Results for the first and second halves of the experiments are presented separately in panels A and B. Error bars represent 1 standard error, and asterisks indicate the significance level of the coefficient's difference from zero (**p*<0.05, ***p*<0.01, ****p*<0.001). F_m, F_n and O denote the manipulated feature, non-manipulated feature, and object, respectively.

Effects of regularity manipulation on feature-based learning

To uncover the mechanisms underlying the observed patterns of learning and choice behavior, we next fit the choice behavior of individual participants to multiple reinforcement learning (RL) models. Our goals were to determine whether participants adopted feature-based or object-based learning. In addition, we investigated whether the effects of regularity manipulation were mediated through attention's influences on learning, decision making, or both. To that end, we constructed four new feature-based learning models: one with equal attention on both features (F) and three with attention modulating choice (F_{AC}), learning (F_{AL}), or both (F_{ACL}). Additionally, we included an existing object-based learning model (O) for comparison. We assumed that the effect of regularity manipulation remained constant throughout the session and modeled this effect as a free parameter resulting in separate inverse temperatures for decision making and separate learning rates for the two features, or both (see Computational models in **Methods** for more details). Through model and parameter recovery analyses, we confirmed that these models could be distinguished from each other, and that their parameters were reliably identifiable (**Fig. S1**).

Using the widely applicable information criterion (WAIC) as the goodness-of-fit, we found that participants' behavior was best described by the feature-based learning model with attention modulating only learning (**Fig. 4A**). Interestingly, the model with attention modulating both choice and learning did not improve the fit of data (difference between WAIC of this model and the best model, $\Delta WAIC = 0.93 \pm 2.29$). Moreover, the model with attention modulating only decision making showed a reduced goodness-of-fit compared to the best model ($\Delta WAIC = 11.77 \pm 5.68$), and a similar reduction in fit was observed when removing all attentional modulations ($\Delta WAIC = 34.89 \pm 11.79$). The object-based model had the worst fit overall ($\Delta WAIC = 109.60 \pm 35.10$).

We also examined the stacking weights, which indicate how different models can be combined to maximize predictive performance (Yao et al., 2018). This approach does not assume that the true model is among the included models and can identify models that, despite having a worse overall WAIC, are still useful for predicting behavior. Indeed, we found significant weights not only on the best fitting feature-based learning models (the F_{AL} and F_{ACL} models) but also on the object-based learning model, with stacking weights of 0.38, 0.31, and 0.30 respectively (**Fig. 4B**). Additional control analyses using simplified versions of the F_{AL} model, which had fewer parameters, resulted in a worse fit compared to the full F_{AL} model. This demonstrates the importance of having four separate learning rates and highlights the distinct effects of our manipulation on each learning rate (**Fig. S2**). Overall, results using the fit of choice data showed that participants employed a combination of feature-based and object-based learning strategies, with attention modulating feature-based learning rather than decision making.

To uncover the mechanisms through which regularity modulation influences attention, we next compared the estimated learning rates for the two features using the best-fitting model (**Fig. 4C–F**). We found that the learning rates for the manipulated feature were generally higher than those for the non-manipulated feature, however, the most pronounced difference was observed in the learning rate on the forgone option when it was not rewarded (*median*=0.061*, probability of direction*=0.97, 95% *HPDI:* [-0.0013, 0.13]; **Fig. 4G–J**) (see Parameter estimation and hypothesis testing in **Methods** for more details). This finding further supports the results from mixed-effect regression analysis on choice that showed the effect of attention (evoked by regularity manipulation) on learning is specific to learning about the forgone option (**Fig. 3**).

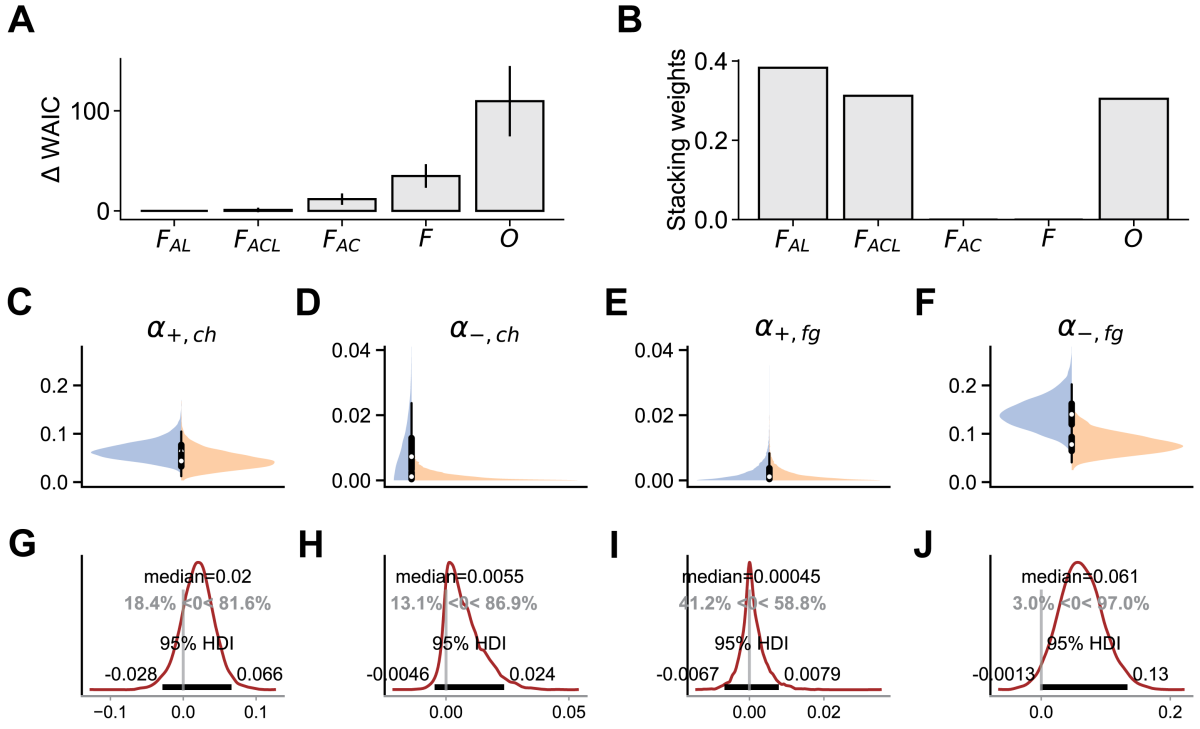


Figure 4. Fitting choice data with different models reveals mechanisms underlying feature-based attentional modulations. (**A**) Comparison of the goodness-of-fit between different models, measured by the difference in WAIC between each model and the best-fitting model, F_{AL} (feature-based learning with attention modulating learning). Model abbreviations: F_{ACL} (feature-based learning with attention modulating choice and learning), F_{AC} (feature-based learning with attention modulating choice), F (feature-based learning with no attention), O (object-based learning). Error bars show 1 s.e.m. (**B**) Plot shows the stacking weights calculated using WAIC. Feature-based learning models that include attention modulating learning have larger weights than models without this feature. Object-based learning also has significant weight despite having overall worse fit. (**C–F**) Evidence of strong confirmation bias in feature-based learning. Plots show posterior distributions of learning rates for chosen (ch) and forgone (fg) options and separately for rewarded (α_+) and unrewarded (α_-) trials. The posterior distributions of the learning rates for the manipulated and non-manipulated features are depicted in blue and orange, respectively. (**G–J**) Evidence of regularity manipulation influencing the learning rates, particularly in unrewarded trials. Plots show the posterior distribution of the differences in learning rates between the manipulated and non-manipulated features, as indicated in panels C–F.

Enhanced confirmation bias due to regularity manipulation

In addition to the differences in learning rates between the manipulated and non-manipulated features, we also observed an overall difference between the learning rates for rewarded and

unrewarded trials. Specifically, the learning rates for the chosen option were higher in rewarded trials than in unrewarded trials (manipulated feature: *median*=0.055*, probability of direction*>0.999, 95% *HPDI*: [0.018, 0.099]; non-manipulated feature: *median*=0.042, *probability of direction*>0.999, 95% *HPDI*: [0.011, 0.083]; **Fig. 4C-D**), a phenomenon referred to as the positivity bias in reward learning (Palminteri & Lebreton, 2022). In contrast, the learning rates for the forgone option were larger in unrewarded than in rewarded trials (manipulated feature: *median*=-0.14, *probability of direction*>0.999, 95% HPDI: [-0.20, -0.075]; non-manipulated feature: *median*=-0.076*, probability of direction*>0.999, 95% *HDPI*: [-0.12, -0.040]; **Fig. 4E-F**), a pattern termed the negativity bias in reward learning.

We found that both types of confirmation bias—positivity and negativity biases—were more pronounced for the manipulated than for the non-manipulated feature. However, this difference was significant for the negativity bias (*median difference* =0.061*, probability of direction*=0.97, 95% *HPDI*: [-0.0025, 0.13]) but not the positivity bias (*median difference*=0.012, *probability of direction*=0.71, *95% HPDI*: [-0.039, 0.059]). Additionally, we found that the negativity bias was larger than the positivity bias but this effect was only significant for the manipulated feature (manipulated: *median difference*=0.083, *probability of direction*=0.98, 95% *HPDI*: [0.013, 0.16]; non-manipulated: *median difference*=0.034, *probability of direction*=0.88, 95% *HPDI*: [-0.024, 0.094]).

Overall, these findings are consistent with the confirmation bias observed in reward learning (Palminteri & Lebreton, 2022). More importantly, our results extend confirmation bias to the realm of feature-based learning with a larger negativity bias compared to positivity bias. Additionally, they suggest that attention may enhance the negativity bias, providing further evidence for the importance of attention in reward learning especially in high-dimensional environments (Niv et al., 2015; Wang & Soltani, 2023).

Changes in learning strategy and effects of regularity manipulation over time

To identify possible changes in learning strategy and the effects of regularity manipulation over time, we first analyzed choice data from the initial half of the trials using the same models as in the previous section. We found that the feature-based learning with attention modulating learning model (F_{AL}) provided the best fit for participants' choice behavior in the first half of the

experiment (**Fig. S3A**). Analysis of the stacking weights further revealed that object-based learning was not useful for predicting behavior (**Fig. S3B**), indicating that participants relied more heavily on feature-based learning at the beginning of the experiment. Moreover, confirmation bias was present in the first half of the experiment (**Fig. S3C–F**), and the effects of regularity on learning rates, specifically the learning rate for the unrewarded forgone options, were more pronounced during this period (**Fig. S3G–J**).

To better assess how feature-based and object-based learning strategies competed against each other over time, we examined the trial-wise differences in WAIC between the best-fitting feature-based model (F_{AL}) and the object-based model (**Fig. 5A**). By fitting a linear mixed-effects model, we found that the trial-wise WAIC significantly decreased as with time (b=-0.04, SE=0.01, t(260.5)=-2.82, p=0.005). The intercept was significantly larger than 0, indicating that the feature-based model had significantly better fit at the beginning of the session (b=0.04, SE=0.007, t(7239)=5.11, p<0.001).

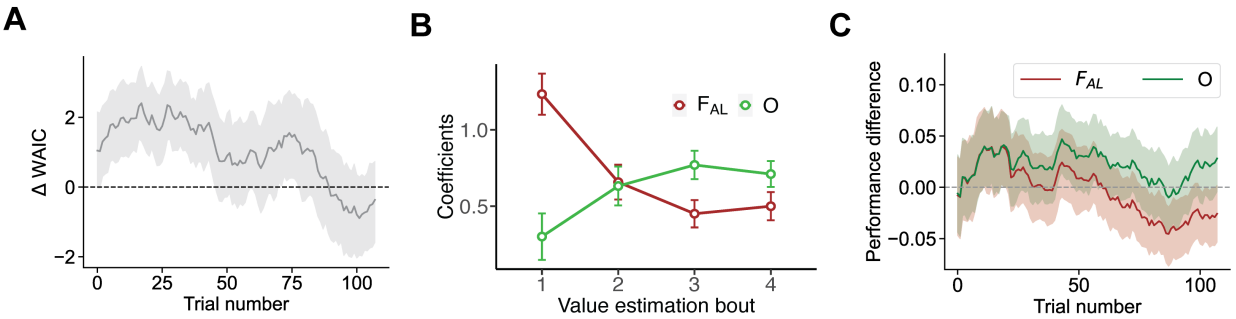


Figure 5. Gradual transition from feature-based to object-based learning. (**A**) Difference in the goodness-of-fit over time. Plot shows the difference in trial-wise WAIC between the best-fitting feature-based learning and object-based learning models. The shaded area represents 1 s.e.m. Initially, the feature-based learning model fits behavior better, but the goodness-of-fit of the two models becomes similar by the end of the experiment. (**B**) Influence of feature-based and object-based learning on value estimates. Plot shows the results of a linear mixed-effects regression predicting value estimates in each bout of estimation trials using reward probabilities from feature-based and object-based models. Each point represents regression coefficients, and error bars represent 1 s.e. Early on, value estimates were more strongly influenced by feature values, but later, feature and object values had similar effects on value estimates. (**C**) Comparison of the learning curves. Plot shows the difference between the participants' learning curves and that of each model. The shaded area represents the 95% highest posterior density interval. Feature-based learning had similar performances as participants but underperformed participants at the end of the session. Object-based learning outperformed participants overall but more closely matched participants' performances in later trials.

In addition, we also used a linear mixed-effects model to predict participants' value estimates during each bout of estimation trials based on the reward probabilities from the feature-based and object-based models (**Table 5**; **Fig. 5B**). This analysis revealed that participants' value estimates were highly feature-based in the first bout of trials. However, later in the experiment, feature and object values had similar influences, with the coefficients on object value being slightly higher. Importantly, this analysis separates the learning of feature and object values from how these values influence participants' value estimates. and thus, demonstrate that participants dynamically arbitrated between feature-based and object-based learning models to control their behavior (Farashahi et al., 2017, 2020; S. W. Lee et al., 2014).

Bout	*Predictor*	*b*	*SE*	*t*	*df*	*p*
1	Intercept	0.37	0.09	4.20	51.95	<0.001
	F_{AL}	1.23	0.13	9.15	49.34	<0.001
	O	0.30	0.15	1.97	63.05	0.053
2	Intercept	0.29	0.07	3.98	78.67	<0.001
	F_{AL}	0.66	0.11	5.74	86.01	<0.001
	O	0.63	0.13	4.97	63.09	<0.001
3	Intercept	0.26	0.05	4.85	89.98	<0.001
	F_{AL}	0.45	0.09	5.01	117.21	<0.001
	O	0.77	0.09	8.22	705.90	<0.001
4	Intercept	0.31	0.06	5.06	75.50	<0.001
	F_{AL}	0.50	0.09	5.42	94.37	<0.001
	O	0.71	0.08	8.37	63.63	<0.001

Table 5. The influences of feature-based and object-based learning systems on participants' value estimates. These influences are estimated by using a mixed-effects logistic regression model to predict participants' value estimates based on reward probabilities from the feature-based with attention modulating learning (F_{AL}) and object-based (O) models.

Finally, we validated the models by simulating the best-fitting feature-based and object-based models using their estimated parameters. We then examined whether the performance of these

models, in terms of the probability of selecting the better option on each trial, matched that of human participants. By calculating the difference between the empirical learning curve and the learning curves of each model, we found that at the beginning of the experiment, both models slightly outperformed our participants (**Fig. 5C**). Soon after, the performance of the feature-based model aligned with that of the participants, whereas the performance of the object-based model remained above that of the participants. However, toward the end of the experiment, the performance of the feature-based model fell below that of the participants, while the performance of the object-based model remained closer to that of the participants. We found consistent results when we conducted similar analyses separately for participants whose choice behavior was better explained by feature-based or object-based learning models (**Fig. S4**).

Together, these results demonstrate that, consistent with previous studies (Farashahi et al., 2017, 2020; Wang & Soltani, 2023), participants dynamically transitioned from learning about features to learning about objects. Moreover, confirmation bias in learning about features became evident early in the experiment when the effects of regularity manipulation were more pronounced. This suggests that attention, evoked by statistical learning, can significantly influence confirmation bias, underscoring the impact of statistical learning on reward learning beyond mitigating the curse of dimensionality.

Discussion

Using a novel experimental paradigm involving two orthogonal tasks combined with multiple computational approaches, we investigated the impact of environmental regularities on reward learning. We found that human participants leveraged regularities about features from a set of stimuli presented during a sequence-prediction task to adjust their learning regarding another set of stimuli that contain similar features. More specifically, both the choices and estimations made by participants during the learning task were more heavily influenced by the objective values associated with the feature with regularity (manipulated feature). This occurred despite the regularities not being informative about rewards. Furthermore, through multiple model-free and model-based analyses, we established that these effects are more consistent with attentional modulations of learning—rather than decision making—as a result of regularity manipulation. Specifically, the learning rates for the feature with regularity were higher, particularly when

learning from the unchosen/forgone option. This suggests that environmental regularities can intensify the confirmation bias observed in reward learning.

Overall, our results demonstrate that learning the regularity about a feature through statistical learning can significantly influence feature-based reward learning and help mitigate the curse of dimensionality. This novel interaction between statistical learning and reward-based learning has interesting implications for learning behaviors, particularly in naturalistic settings. More specifically, in naturalistic settings where choice options possess multiple features, this mechanism could selectively enhance learning of certain features while suppressing others, thereby effectively reducing the dimensionality of the learning task. Similar interaction between reward learning and learning transitions between different states of the environment through successor representation has been shown to offer an efficient learning strategy between model-free and model-based reinforcement learning (Gershman, 2018; Momennejad et al., 2017).

We found that the learning rates were higher for the feature with regularity, possibly due to feature-based attention driven by statistical learning. Importantly, previous studies demonstrated that attentional mechanisms are both used and necessary for statistical learning (Baker et al., 2004; Duncan et al., 2024), and that environmental regularities can modulate attention spontaneously, regardless of goal relevance or intrinsic salience (Zhao et al., 2013). These attentional modulations can influence various cognitive processes including learning and decision making. In our task, regularities likely directed feature-based attention, which in turn could enhance reward learning. Therefore, the difference we observed in learning rates might be mediated by attentional effects on reward learning, albeit with distinctions from effects reported previously. Specifically, existing research shows that learning and decision making are modulated by attentional weights influenced by the value of features (Leong et al., 2017; Wang & Soltani, 2023). In contrast, we observed that features marked by regularities—not value—biased reward estimations, choice strategies, and learning. Therefore, knowledge acquired through both reward learning and statistical learning can shape future learning.

Interestingly, previous research suggests that both reward value and regularities can modulate the priority map for attentional selection (Kim & Anderson, 2021; Le Pelley et al., 2022; Ogden et al., 2023; Stankevich & Geng, 2014). For instance, an object previously associated with a reward can increase the reaction time in a visual search task, even if it is irrelevant to the task (Anderson

et al., 2011). This indicates that reward value can inherently drive attentional mechanisms, regardless of task relevance. As mentioned earlier, reward learning is also influenced by value-driven attention (Leong et al., 2017; Wang & Soltani, 2023). In our study, we controlled for the informativeness of two features to ensure that any observed differences in feature-based learning were due to regularities rather than reward value. Furthermore, to prevent contamination of reward value from the learning task influencing feature-based selection in the sequence-prediction task, we employed two completely different sets of stimuli for each task.

We found that attention evoked by environmental regularities primarily affects learning, rather than decision making. In contrast, prior studies on value-based attention have shown its impact on both learning and decision making (Leong et al., 2017; Niv et al., 2015), whereas others suggested it influences only the learning process, not decision making (Wang & Soltani, 2023). This difference may stem from variations in the experimental paradigms used or from the inherent challenges in detecting attentional modulation within RL models. Most previous studies used models in which only the value of the chosen option was updated based on reward feedback while allowing the value of the unchosen option to decay. In our study, however, feedback was presented for both the chosen and unchosen options allowing us to capture the differential effect of attention on updating the values of both chosen and unchosen options. This could favor the specific effect of attention during learning we found. However, our model recovery results suggest a bias towards detecting attentional effects on learning rather than on decision making. This indicates that there could be undetected effects of attention on choice.

Our finding that regularity manipulation strongly influences reward learning through attentional mechanisms suggests that attention driven by statistical learning could be as powerful, if not more so, in shaping behavior than attention driven by reward values. This is consistent with the literature on effects of reward on perception and working memory (Hickey et al., 2010b, 2010a, 2014; Infanti et al., 2015) and statistical learning literature (Jiang et al., 2015; Rogers et al., 2016). In a study comparing the two types of effects, Jiang et al. (2015) used a task in which participants searched for a ‘T’ among ‘L’ letters while their attention was directed to one of the visual quadrants. They found that attention driven by statistical learning significantly affected search times, whereas the impact of reward-driven attention was negligible. Overall, these findings suggest that in multidimensional environments, the statistical properties of features may

exert a stronger influence on reward learning than the reward information carried by the features themselves. Although this statistical learning helps mitigate the curse of dimensionality, it could also introduce biases that diminish the accuracy of reward learning.

Studies of confirmation bias in reinforcement learning have shown a preference for learning from confirmatory rather than disconfirmatory evidence in both human and non-human animals (Doll et al., 2011; Lefebvre et al., 2022; Palminteri et al., 2017). This bias has significant implications for reinforcement learning, potentially leading to both biased and optimal learning outcomes depending on the reward environment (Palminteri & Lebreton, 2022). Here, we find evidence for the presence of similar confirmation bias in feature-based learning, suggesting that its underlying mechanisms are widely applicable. This effect was present early in the experiment and was not affected by transition from feature-based to object-based learning.

Interestingly, we observed a stronger negativity bias than positivity bias, a pattern potentially influenced by our task design in which we provided reward feedback for the forgone option rather than leaving it to be inferred. Nonetheless, this contrast suggests that mechanisms underlying learning from factual versus counterfactual outcomes may be fundamentally different (Boorman et al., 2011; D. Lee et al., 2005). Additionally, our results demonstrate that attention evoked by detecting regularities influences both types of confirmation bias, especially negativity bias, underscoring the role of attention in reward learning beyond mitigating the curse of dimensionality.

Given the observed effects of regularity on reward learning, one might question whether the incorporation of regularities ultimately benefits learning or could lead to maladaptive behavior. In our study, we ensured that statistical relationships in the environment are not predictive of reward outcomes and therefore, faster learning from the feature with regularity is not beneficial. Similarly, if the feature with regularity is not informative about reward, the impact of regularity incorporation on learning can be maladaptive. In many cases, however, learning the statistical structure of reward outcomes can improve reward learning by providing predictions in addition to enhancing simple reward associations (Atilgan et al., 2022; Fritsche et al., 2023; Gibson, 1969; Klein-Flügge et al., 2019; Soltani & Koechlin, 2022; Woo et al., 2023). For example, during reversal learning, animals can learn and expect the time point at which reversals happen, leading to faster reversals (Costa et al., 2016; Jang et al., 2015; Woo et al., 2023). Moreover,

acquiring statistical knowledge of the environment can facilitate faster learning of new or modified sequences compared to the rigid stimulus-stimulus associations typically seen in purely RL models (Klein-Flügge et al., 2019). Therefore, while detecting features with regularities and using them for learning can mitigate the curse of dimensionality, these regularities could also carry information about environmental changes related to finding food (e.g., day-night cycles), shelter (e.g., seasonal changes), mates (e.g., hormonal cycles), etc. Consequently, detecting these regularities and utilizing them for learning can ultimately help survival while facilitating learning in high-dimensional reward environments.

Methods

Participants

78 participants were recruited from Dartmouth College (53 females, age 18-32, mean age = 19.65). We excluded 18 participants who did not perform significantly higher than chance level (50%) in the second half of the choices trials according to a binomial test (α=0.05) or whose value estimations did not correlate significantly with the true values (using Spearman correlation, α=0.05) in the final two estimation bouts. We also excluded two participants who did not perform significantly higher than chance during the sequence-prediction task. Finally, one participant was excluded who failed to complete the experiment due to a hardware malfunction. This led to the inclusion of 57 participants in the analysis. Participants were compensated based on their performance (up to $10) in addition to a baseline compensation (either $10 or 1 T-point which is Dartmouth College credit compensations for some courses within the Psychological and Brain Sciences Department). During the experiment, accumulated reward points were displayed on the screen using a reward bar after each reward feedback, with every 10 reward points earning them 1 dollar. The experiment was approved by Dartmouth college Institutional Review Board, and a written consent form was obtained from each participant at the beginning of the experiment.

Visual stimuli

The experiment was created using jsPsych (Leeuw et al., 2023) and presented within a web browser of the computer in a controlled experiment room. We created the stimuli from a set of

"Quaddles" all in grayscale (Watson et al., 2019). Each stimulus was composed of two features (shape and pattern) from eight unique shapes and eight patterns. For each participant, two distinct sets of 16 stimuli were created from four shapes and four patterns each, ensuring no features overlapped between the two sets. One set was used for the multidimensional learning task, and the other for the sequence-prediction task, with assignments counterbalanced across participants. This separation aimed to eliminate any bias from the effect of temporal regularities in the sequence-prediction task on participants' expectations about upcoming features during the learning task.

Reward schedule and object-based vs. feature-based learning strategy

We designed a specific reward schedule to allow adoption of different learning strategies while ensuring that two features of stimuli/object are equally informative about reward outcomes (**Fig. 1D**). More specifically, an object-based learner chooses between the two choice options (stimuli) by estimating the reward probabilities of the two options directly. In contrast, a feature-based learner estimates the reward probabilities of the features separately and combines them to estimate reward probability for each stimulus/object. The feature values are defined as the average reward probabilities of stimuli/objects that contain a given feature:

$$Pr(F_{i,j}) = \sum_{\forall\, O_k \text{ containing } F_{i,j}} Pr\,(O_k)/4 \qquad \text{(Eq. 1)}$$

where $F_{i,j}$ denotes the j-th instance of the i-th feature, O_k denotes the k-th object, and Pr denotes the probability of reward.

Using the Naive Bayes equation (Farashahi et al., 2017; Hunt et al., 2014), one can combine the feature values to estimate the stimulus/object values:

$$\overline{Pr}(O_k) = \frac{Pr(F_{1,a}) \times Pr(F_{2,b})}{Pr(F_{1,a}) \times Pr(F_{2,b}) + (1 - Pr(F_{1,a})) \times (1 - Pr(F_{2,b}))} \quad \text{where } O_k \text{ contains } F_{1,a} \text{ and } F_{2,b} \qquad \text{(Eq. 2)}$$

Taking the logit transformation of both sides reveals that the log odds of reward for each object can be estimated by linearly combining the log odds of reward for each feature:

$$log\,\frac{\overline{Pr}(O_k)}{1-\overline{Pr}(O_k)} = log\,\frac{Pr(F_{1,a})}{1-Pr(F_{1,a})} + log\,\frac{Pr(F_{2,b})}{1-Pr(F_{2,b})} \quad \text{(Eq. 3)}$$

However, an agent may calculate a weighted combination of feature values to estimate the reward probabilities associated with each stimulus or object:

$$log\,\frac{\overline{Pr}(O_k)}{1-\overline{Pr}(O_k)} = w_1\,log\,\frac{Pr(F_{1,a})}{1-Pr(F_{1,a})} + w_2\,log\,\frac{Pr(F_{2,b})}{1-Pr(F_{2,b})} \quad \text{(Eq. 4)}$$

where w_1 and w_2 are the weights of the two features. Estimation of object values based on weighted average of feature values will differ slightly from the actual reward probabilities of objects, leading to estimation errors that allow us to disentangle feature-based and object-based strategies. In our experiment, each instance of feature 1 had a corresponding instance in feature 2 with the same reward probability, meaning both features were equally informative about the reward outcome (**Fig. 1D**). This symmetric reward schedule was implemented to prevent value-based attentional biases toward any particular feature (Farashahi et al., 2017; Farashahi & Soltani, 2021; Leong et al., 2017; Niv et al., 2015; Wang & Soltani, 2023).

Experimental procedure

The experiment consisted of two tasks and three types of trials: the sequence-prediction task that involved only one type of trials, and the multidimensional learning task that involved choice and estimation trials (**Fig. 1A**). The sequence-prediction task was used to introduce temporal regularity about a specific feature. In each trial of this task, participants were presented with a sequence of visual stimuli and were required to identify the next stimulus in the sequence. Each stimulus consists of two features, a shape, and a pattern. While the sequence of one of the features was predictable (repeating doublets/triplets/quadruplets), the sequence of the other one was presented in a random fashion. Temporal regularity of this kind has been shown to guide feature-based learning towards the more predictable feature (Chun & Jiang, 1999; Zhao et al., 2013), which we will refer to as the manipulated feature. The length of the sequence varied between seven to eleven stimuli/objects. Following the presentation of a sequence, two choice options were presented (**Fig. 1A**). After making a choice, participants received feedback indicating whether it was correct, along with the correct sequence. Each participant completed up to 20 practice sequence-prediction trials at the start of the experiment. After the first five practice

trials, those who achieved five consecutive correct predictions could immediately proceed to the main experiment. During the main experiment, one sequence-prediction trial occurred before every set of eight choice trials.

During choice trials of the learning task, participants had to choose between a pair of stimuli/objects using the F and J keys on a keyboard. After choosing between the two options, a white circle would appear around the chosen stimulus/object indicating the choice of the participant. Feedback for both stimuli were then presented (green circle and +1 for the rewarded and red circle and -1 for unrewarded option). An accumulated reward point representing the points the participant had earned would be shown at the bottom of the screen and after every 10 reward points followed by a message "You earned one dollar!" and then would be reset again. For the choice trials, each stimulus/object was assigned a reward probability unknown to the participants and had to be learned through trial and error. The relationship between the perceptual properties of the stimuli and the reward probability was randomly shuffled across participants. We selected a subset of 32 pairs of stimuli/objects to show during choice trials (**Fig. 1D**). The pairs were selected to lower the difficulty of the task, while still allowing us to detect different learning strategies. There were a total of 128 choice trials in the experiment. Each pair of choice options was presented in varying order arrangements during the first 64 trials and again in the last 64 trials, with the orders randomized across participants.

In value estimation trials of the learning task, we asked participants to indicate their estimation of reward probability for each stimulus/object using a sliding scale (**Fig. 1A**). In each bout of estimation trials, participants were shown each of the 16 stimuli/objects in random order, one in each trial, and were asked to estimate the reward probability of the presented object with a slider ranging between 0-100%. There was no feedback in the estimation trials. One bout of value estimation trials happened after every 32 choice trials.

Critically, we made the sequence-prediction and learning tasks orthogonal in two ways. First, we separated the reward and regularities by not providing reward during the sequence-prediction task. Second, we used different sets of stimuli/objects in the two tasks so that the learning about stimuli/objects in one task does not influence the learning about stimuli/objects in the other task.

Finally, to encourage prompt responses and engagement, we set a time limit of fifteen seconds for the sequence-prediction and value-estimation trials, and eight seconds for choice trials.

Warnings were displayed at ten seconds, seven seconds, and five seconds before the time limit in sequence-prediction trials, value-estimation trials, and choice trials, respectively. After reaching the time limit, the trial was skipped with a warning. The experiment would be terminated if a participant skipped four trials.

Model-free analysis of choice and value estimations

To explore how participants adjusted their choice behavior based on reward and choice history, we fit generalized linear mixed effects models to predict participants trial-by-trial choices using the choice and reward information from the previous trial:

$$\begin{aligned} log\,\frac{p(Choice(t)=L)}{p(Choice(t)=R)} \sim & \sum\nolimits_{i=1}^{2} F_{i,chosen}(t-1) + \sum\nolimits_{i=1}^{2} F_{i,forgone}(t-1) \\ & + O_{chosen}(t-1) + O_{forgone}(t-1) \\ & + \textstyle\sum_{i=1}^{2} R_{chosen}(t-1) \times F_{i,chosen}(t-1) \\ & + \textstyle\sum_{i=1}^{2} R_{forgone}(t-1) \times F_{i,forgone}(t-1) \\ & + R_{chosen} \times O_{chosen}(t-1) + R_{forgone} \times O_{forgone}(t-1) \end{aligned} \qquad \text{(Eq. 5)}$$

$F_{i,chosen}$ is equal to +1 if the left option shared the i-th feature with the previously chosen option, -1 if the right option shared the i-th feature with the previously chosen option, and 0 otherwise. $F_{i,forgone}$ is defined analogously but based on the previously forgone option. O_{chosen} and $O_{forgone}$ are defined analogously depending on whether the choice option (stimulus/object) on a given trial is the same as the stimuli/object on the previous trial. R_{chosen} and $R_{forgone}$ are the rewards associated with the chosen and forgone options (rewarded: +1; unrewarded: -1).

To examine participants' value representation more directly, we fit linear mixed effects models to predict participants' value estimations from the feature and object values. Both value estimations and the feature and object values were put through a logit transformation ($p \rightarrow log\,(\frac{p}{1-p})$), motivated by Eq. 3. The value estimations were clamped between 0.01 and 0.99 to prevent infinite values. The boundaries for clamping did not qualitatively change our results. To capture any variations in value estimations that deviate from the objective feature values, we also

fit linear mixed effects models using the identities of the two features and their interaction as the independent variables.

All mixed effect modeling was performed using the lme4 and lmerTest packages (Bates et al., 2015; Kuznetsova et al., 2017). By default, models with maximal random effect structure were fit. If the model failed to converge, we incrementally simplified the model by first removing the correlation parameters between random effects, and then any random slope or random intercept that had small standard deviations if it prevented convergence and did not significantly improve model fit. Differences between parameters were tested using contrast tests. A t-test was used for testing contrasts in linear mixed-effects models, and a Wald Chi-square test was used for generalized linear mixed-effects models (Fox & Weisberg, 2019).

Computational models

Object-based RL. The object-based model assumes that participants learn the value of each object directly using a Rescorla-Wagner type reinforcement learning (RL) model. Because we provided reward feedback for both the chosen and unchosen/forgone options, we created a model that updates the values of both of the chosen and the unchosen stimuli. Moreover, we used separate learning rates for rewarded or unrewarded trials. This leads to four learning rates and the following update rules:

$$
\begin{aligned}
V_{ch}(t+1) &= V_{ch}(t) + \alpha_{+,ch}\big(1 - V_{ch}(t)\big), && if\ r(t) = 1 \\
V_{ch}(t+1) &= V_{ch}(t) - \alpha_{-,ch}\big(V_{ch}(t)\big), && if\ r(t) = 0 \\
V_{fg}(t+1) &= V_{fg}(t) + \alpha_{+,fg}\left(1 - V_{fg}(t)\right), && if\ r(t) = 1 \\
V_{fg}(t+1) &= V_{fg}(t) - \alpha_{-,fg}\left(V_{fg}(t)\right), && if\ r(t) = 0
\end{aligned}
\qquad \text{(Eq. 6)}
$$

where $V_{ch}(t)$ and $V_{fg}(t)$ denote the values of chosen and forgone options on trial *t*, *r*(*t*) is the reward outcome on trial *t* (1: rewarded, 0: unrewarded), $\alpha_{+,ch}$and $\alpha_{-,ch}$ denote the learning rates for the chosen option on rewarded and unrewarded trials, and $\alpha_{+,fg}$ and $\alpha_{-,fg}$denote the learning rates for the forgone option on the rewarded and unrewarded trials.

We assumed that in each trial, the decision maker chooses between the two options based on the current estimate of reward probabilities of the two options. Hence, the probability of choosing option represented on the left (Pr_{left}), over option represented on the right (Pr_{right}) was calculated as follows:

$$log\ (\frac{Pr_{left}(t)}{Pr_{right}(t)}) = \beta(V_{O,left}(t) - V_{O,right}(t)) + bias \qquad \text{(Eq. 7)}$$

where $V_{O,\text{left}}(t)$ and $V_{O,\text{right}}(t)$ denote the values of the object on the left and right, β is the inverse temperature with lower values (high temperature) resulting in more stochastic choice, and *bias* is the participant's side bias in choosing between the two options.

It has been shown that the reward values of the unchosen object or feature(s) decay over time (Farashahi et al., 2017; Farashahi & Soltani, 2021; Niv et al., 2015). To capture this effect, we introduced a forgetting factor by adding a decay parameter to the RL models. More specifically. the reward probabilities (values) of the objects/features that were not updated due to being unavailable on a given trial decay to 0.5 with a constant rate of *d*:

$$V(t+1) = V(t) - d \times (V(t) - 0.5) \qquad \text{(Eq. 8)}$$

This model has 7 parameters in total, including one bias, an inverse temperature, four learning rates, and one decay parameters.

Feature-based RL. We assume learners who adopt the feature-based learning strategy, learn the predictive values of each feature instance separately and then estimate the value of each stimulus/object by combining these feature values. For this type of learner, we assumed an update rule similar to Eq. 6 but separately for each feature, and potentially different learning rates for the two features. We then used a weighted linear combination the feature values to predict choice as follows:

$$log\ (\frac{Pr_{left}(t)}{Pr_{right}(t)}) = \beta_{shape}(V_{shape,left}(t) - V_{shape,right}(t))$$

$$+\beta_{pattern}(V_{pattern,left}(t) - V_{pattern,right}(t)) + bias \qquad \text{(Eq. 9)}$$

where V_{shape} and $V_{pattern}$ denote the values of feature shape and pattern, β_{shape} and β_{shape} denote the inverse temperatures for the influence of shape and pattern value on choice, and *bias* is the

participant's side bias in choosing between the two options. Because sequence-prediction trials precede choice trials, we assumed that the effect of attention to be constant throughout the entire session. The full model with both separate learning rates and separate decision weights for both features is referred to as the feature-based learning model with attention modulating both learning and decision making.

Model fitting and model selection

We fit the above RL models (and their variants) to the participants' choice behavior. The models were fit to the experimental data using a hierarchical Bayesian approach (Ahn et al., 2017) implemented in NumPyro (Bingham et al., 2019; Phan et al., 2019). The parameters were sampled using the No-U-Turns sampler, an efficient variant of the Markov chain Monte Carlo (MCMC) algorithm. Specifically, four chains were run in parallel and 3000 samples were taken from each chain. The first 1000 from each were used as warm-up samples and discarded from further inference, resulting in a total of 8000 posterior samples. The hierarchical Bayesian approach aggregates information from multiple participants to inform the group-level parameter estimates while also using the group-level parameter estimates to regularize estimates of individual parameters (Ahn et al., 2017), making it a good model fitting methodology for fitting complex cognitive models where each participant goes through a limited number of trials.

Following recommendations by Ahn et al. (2017), a group-level prior of Normal(0, 1) was put on the means of all the parameters except for the bias, which had a prior of Normal(0, 10). A group-level prior of HalfCauchy(0, 10) was put on the standard deviation of all the parameters. The participant-level parameters were reparameterized by shifting and scaling standard normal random variables with the group means and standard deviations. All participant-level learning rates and decay rates were then put through an inverse probit transform to be bound between 0 and 1, and all participant-level inverse temperatures were put through an exponential transform to ensure positivity. For the control analyses with tied learning rates, the effect of attention was modeled as an additional parameter being added to the learning rates of the manipulated feature before the inverse probit transform. It had the same prior as all other learning rate parameters. All parameters had an r-hat less than or equal to 1.01, effective sample size of at least 400, and no model had divergence.

We compare the widely applicable information criterion (WAIC) between different models as a measure of the goodness-of-fit and the stacking weights (Vehtari et al., 2017; Yao et al., 2018). WAIC approximates the expected out-of-sample predictive performance of a model using its in-sample performance and complexity, which includes the effective number of parameters (Vehtari et al., 2017). Stacking weights optimally combine different models to achieve the highest predictive performance (Yao et al., 2018). Models that are more effective in prediction are assigned higher stacking weights. WAIC was calculated using leave-one-subject-out approximate cross validation, with the exception of trial-wise WAIC. We consider an improvement of WAIC bigger than 1.96 times the standard error to be significant (Vehtari et al., 2017). Model comparisons and parameter estimations were computed using the ArviZ package (Kumar et al., 2019). We used 1000 randomly selected posterior samples for model validation simulations, a method referred to as the posterior predictive check (Ahn et al., 2017; Baribault & Collins, 2023). We found that increasing the number of samples did not qualitatively change the results.

Parameter estimation and hypothesis testing

To perform hypothesis testing on the value of a given parameter (Makowski et al., 2019), we report the median of the posterior samples of that parameter as a measure of central tendency. Furthermore, we report the probability of direction, which is the proportion of posterior samples that are positive or negative, whichever is higher. This is analogous to one minus the one-sided p-value. We interpret a probability of direction greater than 0.95 as evidence that a parameter is significantly greater than or smaller than 0. We also report the 95% highest posterior density interval (95% HPDI), which is the interval that contains 95% of the posterior samples, and such that points within the interval have a higher probability density than points outside of it.

All hypothesis testing on parameters were performed using the posterior distribution of the group-level mean parameters unless otherwise noted. We defined positivity and negativity biases as the difference between the learning rates from reward and no reward for either the chosen or forgone options. We defined the effect of regularity manipulation (attention) as the difference between the learning rates for the same choice and reward outcome (chosen and rewarded, unchosen and rewarded, etc.) but separately for different features. Calculations of these effects

were performed for each sample from the posterior distribution to arrive at the posterior distribution of the effect, then the metrics in the previous paragraph were calculated.

Model and parameter recovery

To validate the results of model comparison and parameter estimations, we performed both model recovery and parameter recovery. We used the posterior median of group mean and standard deviation as the true group-level parameters, and sampled from a normal distribution specified by them to obtain the participant-level parameters. We used these parameters to simulate choice sequences using the same stimulus and reward sequences presented to participants in our experiment (N=57). Note that this type of post-hoc recovery analysis might be conservative because it uses model parameters that were obtained from fitting on the same dataset, potentially leading to more similar behavior between models. The individual differences of parameters with small posterior standard deviation or close to the boundary may be hard to identify. We consider a parameter well-recovered if its 95% HPDI includes the true parameter value used for simulation (Baribault & Collins, 2023).

References

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry (Cambridge, Mass.)*, *1*, 24.

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10367–10371. https://doi.org/10.1073/pnas.1104047108

Atilgan, H., Murphy, C. E., Wang, H., Ortega, H. K., Pinto, L., & Kwan, A. C. (2022). *Change point estimation by the mouse medial frontal cortex during probabilistic reward learning* [Preprint]. Neuroscience. https://doi.org/10.1101/2022.05.26.493245

Baker, C. I., Olson, C. R., & Behrmann, M. (2004). Role of Attention and Perceptual Grouping in Visual Statistical Learning. *Psychological Science*, *15*(7), 460–466. https://doi.org/10.1111/j.0956-7976.2004.00702.x

Baribault, B., & Collins, A. G. (2023). Troubleshooting Bayesian cognitive models. *Psychological Methods*.

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.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P. A., Horsfall, P., & Goodman, N. D. (2019). Pyro: Deep Universal Probabilistic Programming. *J. Mach. Learn. Res.*, *20*, 28:1-28:6.

Boorman, E. D., Behrens, T. E., & Rushworth, M. F. (2011). Counterfactual Choice and Learning in a Neural Network Centered on Human Lateral Frontopolar Cortex. *PLOS Biology*, *9*(6), e1001093. https://doi.org/10.1371/journal.pbio.1001093

Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, *10*(4), 360–365.

Costa, V. D., Dal Monte, O., Lucas, D. R., Murray, E. A., & Averbeck, B. B. (2016). Amygdala and Ventral Striatum Make Distinct Contributions to Reinforcement Learning. *Neuron*, *92*(2), 505–517. https://doi.org/10.1016/j.neuron.2016.09.025

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic Genes Predict Individual Differences in Susceptibility to Confirmation Bias. *Journal of Neuroscience*, *31*(16), 6188–6198. https://doi.org/10.1523/JNEUROSCI.6486-10.2011

Duncan, D., Moorselaar, D. van, & Theeuwes, J. (2024). *Visual statistical learning requires attention*. https://doi.org/10.31234/osf.io/6dezr

Farashahi, S., Rowe, K., Aslami, Z., Gobbini, M. I., & Soltani, A. (2018). Influence of learning strategy on response time during complex value-based learning and choice. *PLOS ONE*, *13*(5), e0197263. https://doi.org/10.1371/journal.pone.0197263

Farashahi, S., Rowe, K., Aslami, Z., Lee, D., & Soltani, A. (2017). Feature-based learning improves adaptability without compromising precision. *Nature Communications*, *8*(1), 1768. https://doi.org/10.1038/s41467-017-01874-w

Farashahi, S., & Soltani, A. (2021). Computational mechanisms of distributed value representations and mixed learning strategies. *Nature Communications*, *12*(1), 7191.

Farashahi, S., Xu, J., Wu, S.-W., & Soltani, A. (2020). Learning arbitrary stimulus-reward associations for naturalistic stimuli involves transition from learning about features to learning about objects. *Cognition*, *205*, 104425.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458–467. https://doi.org/10.1037/0278-7393.28.3.458

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fritsche, M., Majumdar, A., Strickland, L., Garcia, S. L., Bogacz, R., & Lak, A. (2023). *Temporal regularities shape perceptual decisions and striatal dopamine signals* (p. 2023.08.18.553829). bioRxiv. https://doi.org/10.1101/2023.08.18.553829

Gershman, S. J. (2018). The Successor Representation: Its Computational Logic and Neural Substrates. *Journal of Neuroscience*, *38*(33), 7193–7200. https://doi.org/10.1523/JNEUROSCI.0151-18.2018

Gibson, E. J. (1969). *Principles of perceptual learning and development*. Appleton-Century-Crofts.

Hickey, C., Chelazzi, L., & Theeuwes, J. (2010a). Reward Changes Salience in Human Vision via the Anterior Cingulate. *Journal of Neuroscience*, *30*(33), 11096–11103. https://doi.org/10.1523/JNEUROSCI.1026-10.2010

Hickey, C., Chelazzi, L., & Theeuwes, J. (2010b). Reward Guides Vision when It's Your Thing: Trait Reward-Seeking in Reward-Mediated Visual Priming. *PLOS ONE*, *5*(11), e14087. https://doi.org/10.1371/journal.pone.0014087

Hickey, C., Chelazzi, L., & Theeuwes, J. (2014). Reward-Priming of Location in Visual Search. *PLOS ONE*, *9*(7), e103372. https://doi.org/10.1371/journal.pone.0103372

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

Infanti, E., Hickey, C., & Turatto, M. (2015). Reward associations impact both iconic and visual working memory. *Vision Research*, *107*, 22–29. https://doi.org/10.1016/j.visres.2014.11.008

Jang, A. I., Costa, V. D., Rudebeck, P. H., Chudasama, Y., Murray, E. A., & Averbeck, B. B. (2015). The Role of Frontal Cortical and Medial-Temporal Lobe Brain Areas in Learning a Bayesian Prior Belief on Reversals. *Journal of Neuroscience*, *35*(33), 11751–11760. https://doi.org/10.1523/JNEUROSCI.1594-15.2015

Jiang, Y. V., Sha, L. Z., & Remington, R. W. (2015). Modulation of spatial attention by goals, statistical learning, and monetary reward. *Attention, Perception, & Psychophysics*, *77*(7), 2189–2206. https://doi.org/10.3758/s13414-015-0952-z

Katahira, K. (2018). The statistical structures of reinforcement learning with asymmetric value updates.

Journal of Mathematical Psychology, *87*, 31–45.

Kim, H., & Anderson, B. A. (2021). Combined influence of valence and statistical learning on the control of attention: Evidence for independent sources of bias. *Cognition*, *208*, 104554. https://doi.org/10.1016/j.cognition.2020.104554

Kim, H., Sul, J. H., Huh, N., Lee, D., & Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *Journal of Neuroscience*, *29*(47), 14701–14712.

Klein-Flügge, M. C., Wittmann, M. K., Shpektor, A., Jensen, D. E. A., & Rushworth, M. F. S. (2019). Multiple associative structures created by reinforcement and incidental statistical learning mechanisms. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-12557-z

Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, *4*(33), 1143. https://doi.org/10.21105/joss.01143

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, *84*(3), 555–579.

Le Pelley, M. E., Ung, R., Mine, C., Most, S. B., Watson, P., Pearson, D., & Theeuwes, J. (2022). Reward learning and statistical learning independently influence attentional priority of salient distractors in visual search. *Attention, Perception, & Psychophysics*, *84*(5), 1446–1459. https://doi.org/10.3758/s13414-021-02426-7

Lee, D., McGreevy, B. P., & Barraclough, D. J. (2005). Learning and decision making in monkeys during a rock–paper–scissors game. *Cognitive Brain Research*, *25*(2), 416–430. https://doi.org/10.1016/j.cogbrainres.2005.07.003

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.

Leeuw, J. R. de, Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments. *Journal of Open Source Software*, *8*(85), 5351. https://doi.org/10.21105/joss.05351

Lefebvre, G., Summerfield, C., & Bogacz, R. (2022). A Normative Account of Confirmation Bias During Reinforcement Learning. *Neural Computation*, *34*(2), 307–337. https://doi.org/10.1162/neco_a_01455

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*(2), 451–463.

Makowski, D., Ben-Shachar, M. S., Chen, S. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *10*, 2767.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), Article 9. https://doi.org/10.1038/s41562-017-0180-8

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. *Journal of Neuroscience*, *35*(21), 8145–8157.

Ogden, A., Kim, H., & Anderson, B. A. (2023). Combined influence of valence and statistical learning on the control of attention II: Evidence from within-domain additivity. *Attention, Perception, & Psychophysics*, *85*(2), 277–283. https://doi.org/10.3758/s13414-022-02622-z

Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*.

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Computational Biology*, *13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable Effects for Flexible and Accelerated

Probabilistic Programming in NumPyro. *arXiv Preprint arXiv:1912.11554*.

Rogers, L. L., Friedman, K. G., & Vickery, T. J. (2016). No Apparent Influence of Reward upon Visual Statistical Learning. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01687

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

Saffran, J. R., & Kirkham, N. Z. (2018). Infant Statistical Learning. *Annual Review of Psychology*, *69*(Volume 69, 2018), 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Soltani, A., & Koechlin, E. (2022). Computational models of adaptive behavior and prefrontal cortex. *Neuropsychopharmacology*, *47*(1), 58–71. https://doi.org/10.1038/s41386-021-01123-1

Stankevich, B. A., & Geng, J. J. (2014). Reward associations and spatial probabilities produce additive effects on attentional selection. *Attention, Perception, & Psychophysics*, *76*(8), 2315–2325. https://doi.org/10.3758/s13414-014-0720-5

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction, 2nd ed* (pp. xxii, 526). The MIT Press.

Trepka, E., Spitmaan, M., Bari, B. A., Costa, V. D., Cohen, J. Y., & Soltani, A. (2021). Entropy-based metrics for predicting choice behavior based on local response to reward. *Nature Communications*, *12*(1), 6567.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural Evidence of Statistical Learning: Efficient Detection of Visual Regularities Without Awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945. https://doi.org/10.1162/jocn.2009.21131

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit Perceptual Anticipation Triggered by Statistical Learning. *Journal of Neuroscience*, *30*(33), 11177–11187. https://doi.org/10.1523/JNEUROSCI.0858-10.2010

Umemoto, A., Scolari, M., Vogel, E. K., & Awh, E. (2010). Statistical learning induces discrete shifts in the allocation of working memory resources. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(6), 1419–1429. https://doi.org/10.1037/a0019324

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432.

Wang, M. C., & Soltani, A. (2023). Contributions of attention to learning in multi-dimensional reward environments. *bioRxiv*, 2023–04.

Watson, M. R., Voloh, B., Naghizadeh, M., & Womelsdorf, T. (2019). Quaddles: A multidimensional 3-D object set with parametrically controlled and customizable features. *Behavior Research Methods*, *51*, 2522–2532.

Witkowski, P. P., Park, S. A., & Boorman, E. D. (2022). Neural mechanisms of credit assignment for inferred relationships in a structured world. *Neuron*, *110*(16), 2680–2690.

Woo, J. H., Aguirre, C. G., Bari, B. A., Tsutsui, K.-I., Grabenhorst, F., Cohen, J. Y., Schultz, W., Izquierdo, A., & Soltani, A. (2023). Mechanisms of adjustments to different types of uncertainty in the reward environment across mice and monkeys. *Cognitive, Affective, & Behavioral Neuroscience*, *23*(3), 600–619. https://doi.org/10.3758/s13415-022-01059-z

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). *Using stacking to average Bayesian predictive distributions (with discussion)*.

Zhao, J., Al-Aidroos, N., & Turk-Browne, N. B. (2013). Attention Is Spontaneously Biased Toward Regularities. *Psychological Science*, *24*(5), 667–677. https://doi.org/10.1177/0956797612460407

Supplementary materials

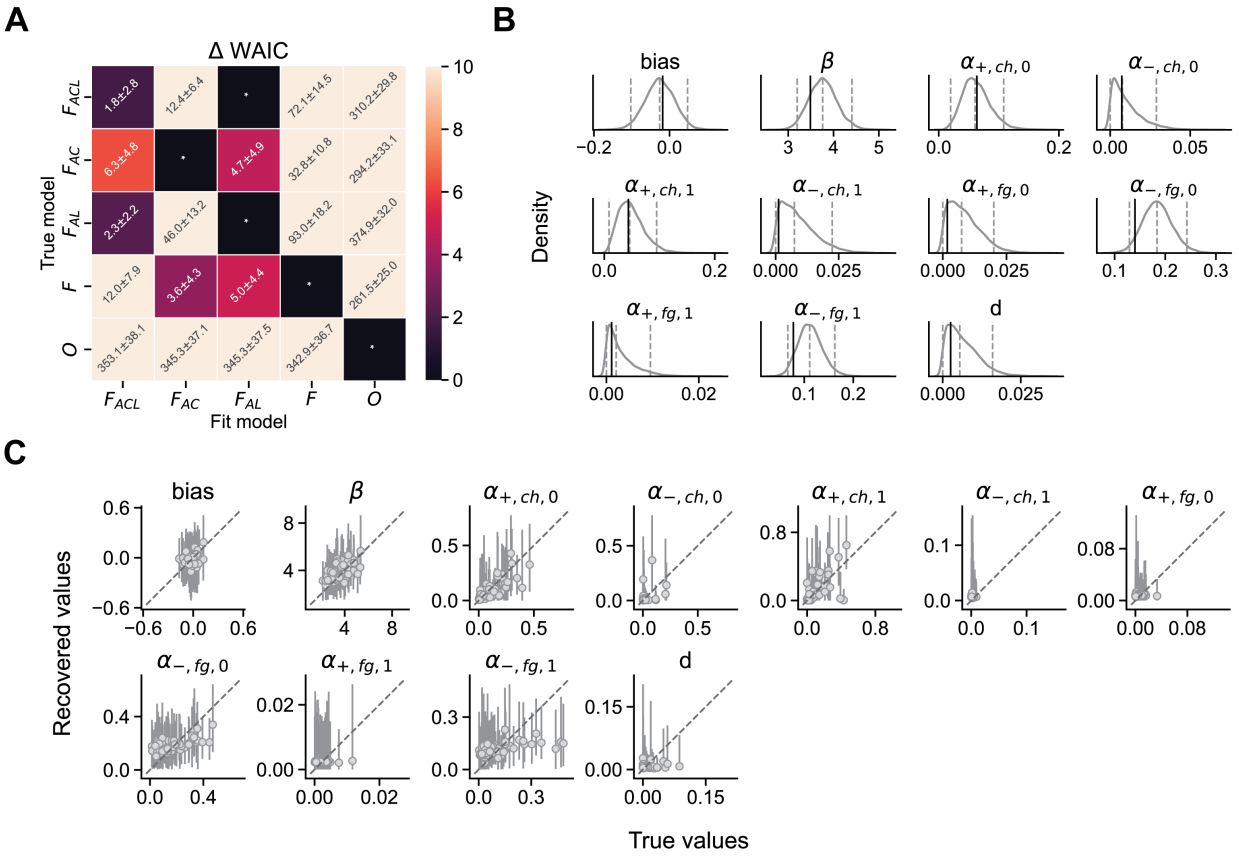


Figure S1. Model and parameter recovery. (**A**) Results of model recovery. Reported are the WAIC difference ± SE between each model and the best-fit model. An asterisk marks the best-fit model, which accurately identifies the correct model in all cases except for the F_{ACL}. Although some feature-based models may be mistaken for each other, no incorrect model provides a significantly better fit than the true model. (**B**) Recovery of group-level mean parameters shows that the 95% HPDI of all parameters includes the true parameter values, indicating an acceptable level of recovery. (**C**) At the participant level, the 95% HPDI of 96.97% of the recovered parameters includes the true parameter values, confirming an overall acceptable level of recovery.

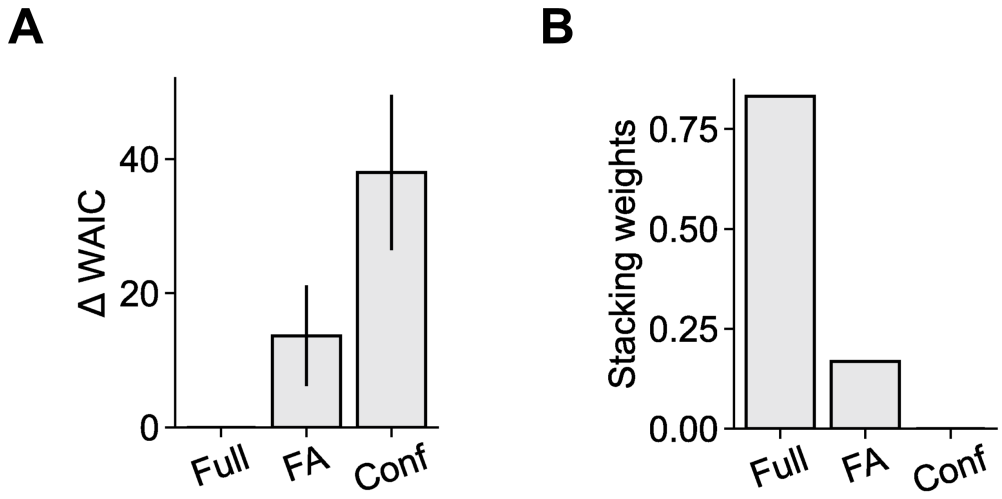


Figure S2. All components of the best model (feature-based learning model with attention modulating learning, F_{AL}) are crucial for capturing choice behavior. (**A**) Comparison of the goodness-of-fit between the full model (F_{AL}) and two variations that lack some of the components, as measured by the difference in WAIC between each model and the full model. Error bars show 1 s.e.m. The full model assumes separate learning rates for each feature, rewarded and unrewarded trials, and for chosen and forgone options (total 8 learning rates). The feature-attention (FA) model simplifies this by assuming separate learning rates for rewarded and unrewarded trials, and for chosen and forgone options (total 4 learning rates). In addition, attention in the FA model is treated as a single multiplicative factor that modulates all 4 learning rates associated with the manipulated feature. The confirmation model (Conf) assumes two separate learning rates for confirming or disconfirming outcomes and for each feature (total 4 learning rates) (Palminteri & Lebreton, 2022). Both simplifications led to worse fit compared with the full model ($\Delta WAIC = 13.65 \pm 7.50$ and 38.01 ± 11.58). (**B**) Plot shows the stacking weights calculated using WAIC. A weighted combination of the full F_{AL} model and the feature-attention (FA) model achieved the highest predictive performance, with the FA model only having a small weight.

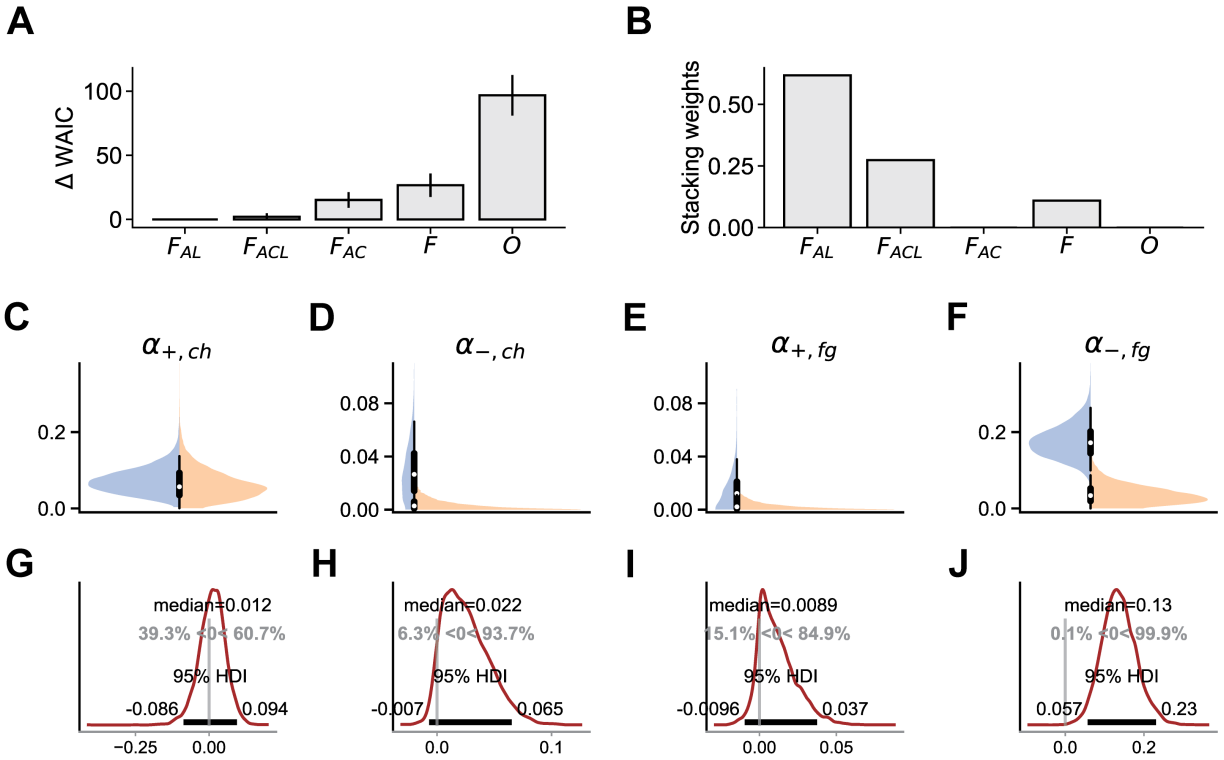
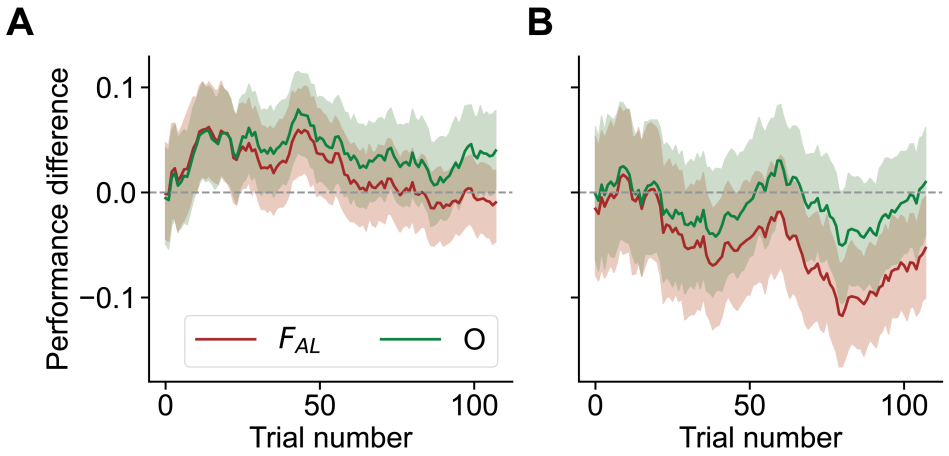


Figure S3. Results of fitting choice data from the first half of trials using different models and estimated parameters based on the best model. Corresponding results for the fit based on all data are shown in Figure 4. (**A**) Comparison of the goodness-of-fit between different models as measured by the difference in WAIC between each model and the best-fitting model (feature-based learning with attention modulating learning, F_{AL}). Error bars show 1 s.e.m. (**B**) Plot shows the stacking weights calculated using WAIC. Feature-based learning models that include attention modulating learning have larger weights than models without this feature and moreover, participants did not utilize object-based learning in the first half of the experiment. (**C–F**) Evidence of strong confirmation bias in feature-based learning. Plots show posterior distributions of learning rates for chosen (ch) and forgone (fg) options and separately for rewarded (α_+) and unrewarded (α_-) trials. The posterior distributions of the learning rates for the manipulated and non-manipulated features are depicted in blue and orange, respectively. (**G–J**) Regularity manipulation influences the learning rates more strongly in the first half of trials, particularly for learning from the forgone option and in unrewarded trials (see Figure 4G–J for comparison). Plots show the posterior distribution of the differences in learning rates between the manipulated and non-manipulated features, as indicated in panels C–F.



Figures S4. Differences between simulated and empirical learning curves, shown separately for participants whose choice behavior was better fit by the feature-based model (F_{AL}, N=38) in panel (A), and the object-based model (O, N=19) in panel (B). The object-based model consistently outperformed participants who adopted the feature-based strategy, but closely matched the performance of participants using the object-based strategy. Conversely, the feature-based model underperformed compared to object-based participants, but more accurately matched the performance of those using the feature-based strategy. Overall, this indicates that between-participant variability in learning strategies (feature-based vs. object-based) partially explains the results shown in Figure 5C.