Novel entropy-based metrics for predicting choice behavior based on local response to reward

Short title: Predicting choice behavior from local response to reward

Ethan Trepka¹⁺, Mehran Spitmaan¹⁺, Bilal A. Bari², Vincent D. Costa³, Jeremiah Y. Cohen²,
Alireza Soltani^{1*}

¹*Department of Psychological and Brain Sciences, Dartmouth College, NH, USA*

²*The Solomon H. Snyder Department of Neuroscience, Brain Science Institute, Kavli Neuroscience Discovery Institute, The Johns Hopkins University School of Medicine, MD, USA*

³*Department of Behavioral Neuroscience, Oregon Health and Science University, OR, USA*

⁺ *These authors contributed equally to this work*

*Corresponding author: AS, Department of Psychological and Brain Sciences, Dartmouth College, Hanover NH 03755, soltani@dartmouth.edu.

Conflict of interest: The authors declare no competing interests.

Manuscript information: 5 figures, 7 supplementary figures, 2 supplementary tables, 50 pages

Abstract: 150 words

Introduction: 690 words

Discussion: 1200 words

Total (not including Method and figure captions): 6600 words

Keywords: Reinforcement learning, information theory, matching law, win-stay, lose-switch.

Acknowledgments: This work is supported by the National Institutes of Health (Grants R01DA047870 to A.S. and R01NS104834 to J.Y.C.).

Abstract

For decades, behavioral scientists have used the matching law to quantify how animals distribute their choices between multiple options in response to reinforcement they receive. More recently, many reinforcement learning (RL) models have been developed to explain choice by integrating reward feedback over time. Despite reasonable success of RL models in capturing choice on a trial-by-trial basis, these models cannot capture variability in matching. To address this, we developed novel metrics based on information theory and applied them to choice data from dynamic learning tasks in mice and monkeys. We found that a single entropy-based metric can explain 50% and 41% of variance in matching in mice and monkeys, respectively. We then used limitations of existing RL models in capturing entropy-based metrics to construct a more accurate model of choice. Together, our novel entropy-based metrics provide a powerful, model-free tool to predict adaptive choice behavior and reveal underlying neural mechanisms.

Introduction

How do we distribute our time and choices between the many options or actions available to us? Around 60 years ago, Richard Herrnstein strived to answer this question based on one of the main ideas of behaviorism; that is, the history of reinforcement is the most important determinant of behavior. He proposed a simple rule called the matching law stating that the proportion of time or responses that an animal allocates to an option or action matches the proportion of reinforcement they receive from those options or actions (Herrnstein, 1961). The matching law has been shown to explain global choice behavior across many species (Williams, 1988) including pigeons (de Villiers and Herrnstein, 1976; William, 1979; Mazur, 1981; Villarreal et al., 2019), mice (Gallistel et al., 2007; Fonseca et al., 2015; Bari et al., 2019), rats (Gallistel, 1994; Belke and Belliveau, 2001; Lee et al., 2017), monkeys (Anderson et al., 2002; Sugrue et al., 2004; Lau and Glimcher, 2005; Kubanek and Snyder, 2015; Tsutsui et al., 2016; Soltani et al., 2021), and humans (Schroeder and Holland, 1969; Pierce and Epling, 1983; Beardsley and McDowell, 1992; Savastano and Fantino, 1994; Vullings and Madelain, 2018; Cero and Falligant, 2020), in a wide range of tasks, including concurrent variable interval, concurrent variable ratio, probabilistic reversal learning, and so forth. A common finding in most studies, however, has been that animals undermatch, corresponding to selection of the better option/action less than it is prescribed by the matching law. Such deviation from matching often corresponds to suboptimal behavior in terms of total harvested rewards, pointing to adaptive mechanisms beyond reward maximization.

The matching law is a global rule but ultimately should emerge from an interaction between choice and learning strategies, resulting in local (in time) adjustments of choice behavior to reinforcement obtained in each trial. Accordingly, many studies have tried to explain matching based on different mechanisms. Reinforcement learning models are particularly useful because they can simulate changes in behavior due to reward feedback. Consequently, recent studies on matching have focused on developing RL models that can generate global matching behavior based on local learning rules (Sugrue et al., 2004; Lau and Glimcher, 2005; Loewenstein and Seung, 2006; Soltani and Wang, 2006; Ito and Doya, 2009; Otto et al., 2011; Iigaya and Fusi, 2013; Saito et al., 2014; Iigaya et al., 2019; Grossman et al., 2020). These RL models are often augmented with some components in addition to stimulus- or action-value functions to improve

fit of choice behavior on a trial-by-trial basis. For example, the models could include learning the reward-independent rate of choosing each option (Lau and Glimcher, 2005), adopting win-stay lose-switch policies (Ito and Doya, 2009; Otto et al., 2011), or learning on multiple timescales (Iigaya et al., 2019). Although these additional components improved the overall fit of choice behavior, the outcome models were not examined for their ability to capture deviation from matching. This could result in misinterpretation or missing important neural mechanisms underlying matching behavior in particular and adaptive behavior more generally (Palminteri et al., 2017; Wilson and Collins, 2019). Therefore, after decades of research on matching behavior, it is still not fully clear how such a fundamental law of behavior emerges as a result of local response to reward feedback.

To address this, we propose a set of novel metrics based on information theory that can summarize trial-by-trial response to reward feedback and predict global matching behavior. To test the utility of our novel metrics, we applied them to large sets of behavioral data in mice and monkeys during two very different dynamic learning tasks. We found that in both mice and monkeys, our entropy-based metrics can predict matching behavior better than existing measures. Moreover, we found a close link between undermatching and the consistency of choice strategy (stay or switch) in response to receiving no reward after selection of the worse option in both species. Finally, we used shortcomings of existing RL models in capturing the pattern of entropy-based metrics in our data to construct a new model that integrates reward- and option-dependent strategies with a reinforcement-learning component. We show that the new model can capture both trial-by-trial choice data and global choice behavior better than the existing models.

Results

Mice and monkeys dynamically adjusted their behavior to changes in reward probabilities.

To study learning and decision making in dynamic reward environments, we examined choice behavior of mice and monkeys during two different probabilistic reversal learning tasks. Mice selected between two actions (licking left and right) that provided reward with different probabilities, and these probabilities changed between blocks of trials without any signal to the animals (**Fig. 1a**; see **Methods** for more details; Bari et al., 2019). Block lengths were drawn from a uniform distribution that spanned a range of 40 to 80 trials. Here, we focused on the

majority (469 out of 528) of sessions in which two sets of reward probabilities (equal to 0.4 and 0.1, and 0.4 and 0.05) were used. We refer to these reward schedules as 40/10 and 40/5 reward schedules (1786 and 1533 blocks with 40/5 and 40/10 reward schedules, respectively). Rewards were “baited” such that if reward was assigned on a given side and that side was not selected, reward would remain on that side until the next time that side was selected. Due to baiting, the

Figure 1. Schematic of the experimental paradigms in mice and monkeys and basic behavioral results. (a–b) Timeline of a single trial during experiments in mice (a) and monkeys (b). To initiate a trial, mice received an olfactory "go" cue (or "no go" cue in 5% of trials) (a), and monkeys fixated on a central point (b). Next, animals chose (via licks for mice and saccades for monkeys) between two options (left or right tubes for mice and circle or square for monkeys) and then received a reward (drop of water and juice for mice and monkeys, respectively) probabilistically based on their choice. (c–d) Average choice and reward using a sliding window with a length of 10 for two representative sessions in mice (c) and monkeys (d). Grey dashed lines indicate trials where reward probabilities reversed. (e–f) Average choice and reward fractions around block switches using a non-causal smoothing kernel with a length of 3 separately for all blocks with a given reward schedule in mice (e) and monkeys (f). The better (or worse) side is the better (or worse) side prior to the block switch. Trial zero is the first trial with the reversed reward probabilities. Average choice fractions for the better side/stimulus are lower than average reward fractions for that side/stimulus throughout the block for both mice and monkeys, corresponding to undermatching behavior.

probability of obtaining reward on the unchosen side increased over time as during foraging in a natural environment. As a result, optimal performance in the task generally required a strategy that involved selecting options in proportion to their reinforcement rates (perfect “matching” behavior). In total, 16 mice performed 469 sessions of the two-probability version of the task for a total of 3319 blocks and 189,199 trials.

In a different experiment, monkeys selected between two new stimuli (a circle or square with variable colors) via saccades and received a juice reward probabilistically (**Fig. 1b**; see **Methods** for more details; Costa et al., 2016). In superblocks of 80 trials, the reward probabilities assigned to each stimulus reversed randomly between trials 30 and 50, such that the more rewarding stimulus became the less rewarding stimulus. We refer to trials before and after a reversal as a block. Monkeys completed multiple superblocks per each session of the experiment wherein the reward probabilities assigned to the better and worse options were equal to 0.8 and 0.2, 0.7 and 0.3, or 0.6 and 0.4, which we refer to as 80/20, 70/30, and 60/40 reward schedules. In contrast to the task used in mice, rewards were not baited. Here, we only analyze data from the 80/20 and 70/30 reward schedules as they provide two levels of reward uncertainty similar to the experiment in mice. In total, 4 monkeys performed 2212 blocks of the task with the 80/20 and 70/30 reward schedules for a total of 88,480 trials.

We found that in response to block switches, both mice and monkeys rapidly adjusted their choice behavior to select the better side or stimulus more often (**Fig. 1c–d**). However, the fraction of times they chose the better side or stimulus fell below predictions made by the matching law, even at the end of the blocks (**Fig. 1e–f**). More specifically, the relative selection of the better side/stimulus (choice fraction) was often lower than the ratio of reward harvested on the better side/stimulus to the overall reward harvested (reward fraction), corresponding to undermatching behavior. Therefore, we next explored how undermatching depends on choice- and reward-dependent strategies.

Mice and monkeys exhibited highly variable undermatching behavior. To better examine matching behavior, we used the difference between the relative choice and reward fractions for each block of trials to define “deviation from matching” (see **Eqs. 1-2** in **Methods**; **Fig. 2a,d**). Based on our definition, negative and positive values for deviation from matching correspond to undermatching and overmatching, respectively. Undermatching occurs when the choice fraction

is smaller than the reward fraction for reward fractions larger than 0.5, or the choice fraction is larger than the reward fraction when the latter is smaller than 0.5. Overmatching occurs when the choice fraction is larger than the reward fraction for reward fractions larger than 0.5, or the choice fraction is smaller than the reward fraction when the latter is smaller than 0.5.

Undermatching could happen because the animal does not detect the more rewarding option/action, poor credit assignment, or due to too much stochasticity in choice. In contrast, overmatching is characterized by selecting the better option more frequently than is prescribed based on perfect matching. In the task used in monkeys, overmatching was not possible by design (except due to random fluctuations in reward assignment) and optimal performance could be achieved by selecting the better stimulus all the time, corresponding to matching. In contrast, while being sub-optimal, overmatching is possible in the reversal learning tasks with baited rewards (e.g., task used in mice).

Consistent with previous studies on matching behavior, we found significant undermatching in mice in both the 40/10 and 40/5 reward schedules (Wilcoxon signed-rank test; 40/10: $Z = -31.2, p = 1.53 \times 10^{-213}$; 40/5: $Z = -35.0, p = 8.75 \times 10^{-269}$; **Fig. 2b–c**). Similarly, we found significant undermatching in monkeys in both the 70/30 and 80/20 reward schedules (Wilcoxon signed-rank test; 70/30: $Z = -27.02, p = 9.74 \times 10^{-161}$; 80/20: $Z = -27.06, p = 3.06 \times 10^{-161}$; **Fig. 2e–f**). In addition, average undermatching for mice was significantly larger in the 40/5 reward schedule than the 40/10 reward schedule, whereas average undermatching for monkeys was not significantly different in the 70/30 and 80/20 schedules (two-sided independent t-test; Mice: $p = 7.93 \times 10^{-36}, d = 0.44$; Monkeys: $p = 0.19, d = 0.06$; **Fig. S1f**). More importantly, undermatching was highly variable in both reward schedules for both mice and monkeys (**Fig. 2b–c, 2e–f**). To understand the nature of this variability, we examined whether existing behavioral metrics and RL models can predict the observed deviation from matching.

Figure 2. Mice and monkeys exhibited highly variable but significant undermatching. (a) Plot shows reward fraction versus choice fraction across all blocks separately for each reward schedule in mice. The black dashed line represents the identity line. For reward fractions greater than 0.5, nearly all points remain below the identity line, indicating choice fraction smaller than reward fraction for the better option (undermatching). Similarly, points above the identity line for reward fractions less than 0.5 indicate undermatching. **(b–c)** Distributions show deviation from matching values for the 40/10 (b) and 40/5 reward schedules (c) in mice. The solid black line indicates 0, corresponding to perfect matching. The dashed black lines are the median deviation from matching. Asterisks indicate significance ($p < 0.01$) based on a two-sided Wilcoxon signed-rank test. **(d–f)** Similar to a–c, but for monkeys with 70/30 and 80/20 reward schedules. Because of random fluctuations in local reward probabilities, overmatching occurred in a minority of cases.

Existing metrics only partially explain variability in undermatching behavior. To examine the relationship between existing behavioral metrics and undermatching, we first performed stepwise multiple regressions to predict deviation from matching for both mice and monkeys based on commonly used metrics including $p(win)$, $p(stay)$, $p(stay|win)$ and $p(switch|lose)$ (threshold for adding a predictor: $p < 0.0001$; see **Methods** for more details). These stepwise regressions resulted in the following equations for predicting deviation from matching in mice and monkeys:

$$Dev. from matching_{Mice} = 0.67 \times p(win) + 0.19 \times p(stay) + 0.12 \times p(stay|win) - 0.65.$$

$$\text{Dev. from matching}_{\text{Monkeys}} = 0.23 \times p(\text{win}) - 0.12 \times p(\text{switch}|\text{lose}) + 0.12 \times p(\text{stay}|\text{win}) - 0.33.$$

These regression models explained 31% and 34% of the variance in deviation from matching for mice and monkeys, respectively, which are significant but unsurprising amounts of overall variance (Mice: $\text{Adjusted } R^2 = 0.31$; Monkeys: $\text{Adjusted } R^2 = 0.34$). In the final regression model, $p(\text{win})$ and $p(\text{stay}|\text{win})$ all had positive regression coefficients for both mice and monkeys, indicating that increased $p(\text{win})$ and $p(\text{stay}|\text{win})$ are associated with less undermatching behavior (note that deviation from matching is mostly negative).

We next included the Repetition Index (RI) on the better (RI_B) and worse (RI_W) options (side or stimulus), which measure the tendency to stay beyond chance on the better and worse options, to predict undermatching (Soltani et al., 2013). To that end, we conducted additional stepwise multiple regressions that predicted deviation from matching using RI_B , RI_W , $p(\text{win})$, $p(\text{stay})$, $p(\text{stay}|\text{win})$, and $p(\text{switch}|\text{lose})$ as predictors. The final regression equations for predicting deviation from matching for mice and monkeys were:

$$\text{Dev. from matching}_{\text{Mice}} = 0.42 \times p(\text{win}) + 0.45 \times p(\text{stay}) + 0.17 \times p(\text{stay}|\text{win}) - 0.44 \times RI_B - 0.75 \times RI_W - 0.73.$$

$$\text{Dev. from matching}_{\text{Monkeys}} = 0.32 \times p(\text{stay}) - 0.57 \times RI_B - 0.04 \times p(\text{switch}|\text{lose}) - 0.33.$$

These models explained 48% and 49% of the variance in deviation from matching for mice and monkeys, respectively (Mice: $\text{Adjusted } R^2 = 0.48$; Monkeys: $\text{Adjusted } R^2 = 0.49$). Thus, including RI_B and RI_W enabled us to account for 17% and 15% more variance, suggesting that staying beyond chance on both the better and worse sides is a significant contributor to undermatching behavior. RI_B and RI_W both had negative coefficients in the regression equations they were present in, indicating that repeating better or worse options beyond chance increases undermatching. This was expected for RI_W because staying beyond chance on the worse side results in more frequent selection of the worse option and thus more undermatching. In contrast, larger RI_B could increase undermatching because more staying on the better side could stop the animals from switching to the new better side after block switches.

Together, these results illustrate that undermatching is correlated with tendency to stay beyond chance (measured by RI) and response to reward feedback in terms of stay or switch (measured by win-stay and lose-switch). However, win-stay and lose-switch are not strong predictors of undermatching because their relative importance depends on the overall probability of winning. For example, if overall win rate is high, lose-switch is less useful for predicting behavior because response to loss represents strategy in a small subset of trials. Although win-stay, lose-switch, $p(\text{win})$, and $p(\text{stay})$ contain all the information necessary to compute the dependence of strategy on reward, this requires interpretation of all four metrics in conjunction and may depend on nonlinear relationships that are challenging to intuit or capture with regression. To overcome these issues, we propose new metrics to quantify changes in strategy due to reward outcome using information theory.

Novel behavioral metrics based on information theory. To better capture the dependence of staying (or similarly switching) strategy on reward outcome, we developed a series of model-free behavioral metrics based on Shannon's information entropy (Shannon, 1948). The Shannon information entropy of a random variable X conditioned on Y , denoted $H(X|Y)$, captures the surprise or uncertainty of X given knowledge of the values of Y . Lower information entropies correspond with decreased uncertainty in the variable under consideration and thus consistency in utilized strategy (see below).

First, we define the entropy of reward-dependent strategy (ERDS) that measures the dependence of staying and switching strategy on reward. Formally, ERDS is the information entropy of strategy conditioned on reward, $H(\text{str}|\text{rew})$ (see **Methods** for more details). ERDS quantifies the amount of information needed to explain an animal's strategy given knowledge of whether the animal won or lost in the previous trial. Lower ERDS values indicate more consistent response to reward feedback.

In its simplest formulation for measuring the effect of reward in the previous trial on staying or switching, ERDS is a function of win-stay, lose-switch, and $p(\text{win})$ (see Eq. 9 in **Methods**). As win-stay and lose-switch move further from 0.5, ERDS decreases, reflecting increased consistency of reward-dependent strategy (**Fig. 3a**). Moreover, $p(\text{win})$ modulates the effects of win-stay and lose-switch on ERDS (**Fig. 3b–c**). As $p(\text{win})$ decreases, the influence of win-stay on ERDS decreases, reflecting that win-stay is less relevant to overall response to reward

feedback when winning rarely occurs. Similarly, as $p(\text{lose}) (=1-p(\text{win}))$ decreases, the influence of lose-switch on ERDS decreases, reflecting that lose-switch is less relevant to response to reward feedback. Because of these properties, ERDS corrects for the limitations of win-stay and lose-switch. Also, as staying (or switching) strategy becomes more independent of reward outcome, ERDS increases because reward outcome provides no information about strategy.

ERDS can be decomposed into ERDS_+ and ERDS_- to measure the specific effects of winning and losing in the preceding trial, respectively (**Fig. 3b–c**; see **Methods**). More specifically, ERDS_+ is the entropy of win-dependent strategy, and ERDS_- is the entropy of loss-dependent strategy. Therefore, comparing ERDS_+ and ERDS_- provides information about the relative contributions of win-dependent strategy and loss-dependent strategy to overall reward-dependent strategy.

In addition to conditioning stay strategy on reward in the preceding trial, we can also condition stay strategy on selection of the better or worse option. The resulting entropy of option-dependent strategy (EODS), $H(\text{str}|\text{opt})$, captures the dependence of staying and switching strategy on the selection of the better or worse option (or action) in the preceding trial (**Fig. 3d**). EODS depends on $p(\text{choose better})$, $p(\text{stay}|\text{choose better})$, and $p(\text{switch}|\text{choose worse})$ and moreover, can be decomposed into EODS_B and EODS_W based on selection of the better or worse option (or action), respectively (**Fig. 3e–f**).

Finally, to capture the dependence of strategy on reward and option in a single metric, we computed the entropy of reward- and option-dependent strategy (ERODS), $H(\text{str}|\text{rew}, \text{opt})$. ERODS depends on the probabilities of staying and switching conditioned on all combinations of reward and option (see **Methods** for more details). ERODS has similar properties to ERDS and EODS and can be interpreted in a similar fashion. Low ERODS values indicate that staying and switching strategy consistently depends on combinations of reward and the option selected in the previous trial; for example, winning and choosing the better side in the previous trial. ERODS can be decomposed either by better and worse option (ERODS_B and ERODS_W), by win or loss (ERODS_+ and ERODS_-), or by both (ERODS_{B+} , ERODS_{W+} , ERODS_{B-} , and ERODS_{W-}).

Figure 3. Relationship between new entropy-based metrics and win-stay, lose-switch strategies. (a–c) Plotted are ERDS and ERDS decompositions as a function of $p(\text{win})$, $p(\text{lose})$, win-stay, and lose-switch. Darker colors correspond to larger values of metrics. For the plot in panel A, $p(\text{win})$ is set to 0.5. Observed entropy-based metrics and constituent probabilities for each block for mice (orange dots) and monkeys (green dots) are superimposed on surfaces. (d–f) EODS and EODS decompositions as a function of the probabilities of choosing the better and worse option, $p(\text{better})$ and $p(\text{worse})$, probability of stay on the better option, and probability of switch from the worse option. For the plot in panel (d), $p(\text{better})$ is set to 0.5. For all plots, the units of entropy-based metrics are bits. (g–i) Same as in panels (a–c) but using heatmap. (j–l) Same as in panels (d–f) but using heatmap.

To summarize, we propose three novel metrics, ERDS, EODS, and ERODS that capture the dependence of stay/switch strategy on reward outcome and/or selected option in the preceding trial. Each metric can be decomposed into components that provide important information about the dependence of strategy on winning or losing and/or choosing the better or worse option in the preceding trial (see **Table S1** for summary). We next show how these entropy-based metrics can predict deviation from matching behavior and further be used to construct more successful RL models.

Deviation from matching was highly correlated with entropy-based metrics. To test the relationship between the observed undermatching and our novel entropy-based metrics, we next computed correlations between all behavioral metrics and deviation from matching (**Fig. 4**). We found that nearly all entropy-based metrics were significantly correlated with deviation from matching. Out of all behavioral metrics tested, deviation from matching showed the strongest correlation with $ERODS_w$. based on both parametric (*Pearson*; *mice*: $r = -0.71, p < 10^{-300}$; *monkeys*: $r = -0.64, p = 10^{-231}$) and non-parametric (*Spearman*; *mice*: $r = -0.78, p < 10^{-300}$; *monkeys*: $r = -0.75, p < 10^{-300}$) tests in both mice and monkeys. The size of the correlation between $ERODS_w$ and deviation from matching is remarkable because it indicates that a single metric can capture more than 50% and 41% of the variance in deviation from matching in mice and monkeys, respectively. This finding suggests that undermatching occurs when animals lose when selecting the worse option and respond inconsistently to those losses.

In addition to $ERODS_w$, $EODS_w$ was also highly correlated with deviation from matching (*Pearson*: *mice*: $r = -0.60, p < 10^{-300}$; *monkeys*: $r = -0.53, p = 9.35 \times 10^{-150}$; *Spearman*: *mice*: $r = -0.67, p < 10^{-300}$, *monkeys*: $r = -0.67, p = 1.18 \times 10^{-265}$) as was ERDS. (*Pearson*: *mice*: $r = -0.43, p = 6.72 \times 10^{-156}$; *monkeys*: $r = -0.57, p = 1.94 \times 10^{-191}$; *Spearman*: *mice*: $r = -0.47, p = 7.31 \times 10^{-198}$; *monkeys*: $r = -0.63, p = 1.06 \times 10^{-240}$). Overall, these results show that global deviation from matching was most strongly correlated with the consistency of response after selection of the worse option (or action) and when no reward was obtained.

Figure 4. Correlation between undermatching and proposed entropy-based metrics and underlying probabilities. (a–b) Correlation matrix for 19 behavioral metrics and undermatching in mice using Pearson (a) and Spearman (b) tests. Correlation coefficients are computed across all blocks, and matrix elements with non-significant values ($p > .0001$) are not shown (cells in black). The red rectangles highlight correlation coefficients between behavioral metrics and undermatching. **(c–d)** Similar to (a–b) but for monkeys. Overall, the new entropy-based metrics show stronger correlation with undermatching than previous metrics, and undermatching was most strongly correlated with ERODS_w, EODS_w, and ERDS.

As expected, there were significant correlations between the proposed entropy-based metrics (see lower right of matrices in **Fig. 4**). For instance, ERDS and EODS were highly correlated (*Pearson: mice: $r = 0.90$, $p < 10^{-300}$; monkeys: $r = 0.94$, $p < 10^{-300}$). EODS and ERODS were also highly correlated as expected (*Pearson: mice: $r = 0.95$, $p < 10^{-300}$; monkeys: $r = 0.94$, $p < 10^{-300}$).**

10^{-300} ; *monkeys*: $r = 0.97$, $p < 10^{-300}$). Many entropy-based metric decompositions had similarly large correlations with other entropy-based metric decompositions. Finally, we found consistent results for the relationships between undermatching and our metrics for both reward schedules, even though undermatching and our metrics were sensitive to reward probabilities on the two options (see **Fig. S1**, **Fig. S2**, **Fig. S3**, and **Fig. S4**).

Entropy-based metrics can accurately predict deviation from matching. To verify the utility of entropy-based metrics in predicting deviation from matching, we performed additional stepwise regressions using our entropy-based metrics. In these models, we included $ERDS_+$, $ERDS_-$, $EODS_B$, $EODS_W$, $ERODS_{B+}$, $ERODS_{B-}$, $ERODS_{W+}$, $ERODS_{W-}$, RI_B , RI_W , $p(win)$, $p(stay)$, $win-stay$, and $lose-switch$ as predictors.

The final regression equations for predicting undermatching using all metrics in mice and monkeys were as follows:

$$\begin{aligned} Dev. from matching_{Mice} = & -0.85 \times ERODS_{W-} + 0.32 \times ERODS_{W+} - 0.22 \times ERODS_{B+} + \\ & 0.22 \times ERDS_+ + 0.78 \times EODS_W - 0.37 \times P(switch|lose) - 1.09 \times RI_W + \\ & 0.33 \times P(stay|win) + 0.57 \times ERODS_{B-} - 0.17 \times ERDS_- + 0.11 \times p(win) - \\ & 0.22 \times EODS_B - 0.30 \end{aligned}$$

$$\begin{aligned} Dev. from matching_{Monkeys} = & -1.02 \times ERODS_{W-} + 0.16 \times ERDS_+ + 0.33 \times \\ & P(stay|win) - 0.21 \times P(switch|lose) - 0.66 \times RI_B - 0.23 \times ERODS_{B+} + \\ & 0.83 \times EODS_W - 0.31. \end{aligned}$$

These models explained 74% and 57% of total variance in deviation from matching for mice and monkeys, respectively (Mice: $Adjusted R^2 = 0.74$; Monkeys: $Adjusted R^2 = 0.57$). For mice, the regression model explained 26% more variance than the model with repetition indices and 43% more variance than the model with basic behavioral metrics. For monkeys, the regression model explained 8% more variance than the model with repetition indices and 23% more variance than the model with basic behavioral metrics. These are significant improvements over previous models, suggesting that most variance in undermatching behavior can be explained by trial-by-trial response to reward feedback. Moreover, for mice, the first three predictors added to the regression models were $ERODS_{W-}$ ($\Delta R^2 = 0.59$), $ERODS_{W+}$ ($\Delta R^2 = 0.04$), and $ERODS_{B+}$ ($\Delta R^2 = 0.02$). For monkeys, the first three predictors added were $ERODS_{W-}$ ($\Delta R^2 = 0.31$),

$ERODS_W$ ($\Delta R^2 = 0.09$), and $ERODS_{B+}$ ($\Delta R^2 = 0.06$). Notably, the first three predictors added to the regression models for both mice and monkeys were entropy-based metrics, indicating that entropy-based metrics were the best predictors of deviation from matching when considering all metrics together.

The coefficients of predictors in this model cannot be interpreted in isolation in this model due to multicollinearity among entropy-based metrics. However, seven out of eight and four out of eight of the entropy-based metrics included as predictors were added to the final regression equations for mice and monkeys, respectively. This suggests that despite their overlap, each entropy-based metric captures a unique aspect of the variance in deviation from matching behavior.

Given the complexity of the final equation to predict deviation from matching, we also constructed simpler linear regression models predicting deviation from matching using the first three entropy-based metrics added to the stepwise regressions ($ERODS_{W-}$, $ERODS_{W+}$, and $ERODS_{B+}$ for mice, and $ERODS_{W-}$, $ERODS_{B+}$, and $EODS_W$ for monkeys). For mice, the regression equation for this simple model was: *Dev. from matching* = $-0.62 \times ERODS_{W-} + 0.73 \times ERODS_{W+} - 0.16 \times ERODS_{B+} - 0.02$. For monkeys, the regression equation for this simple model was: *Dev. from matching* = $-1.09 \times ERODS_{W-} - 0.14 \times ERODS_{B+} + 0.63 \times EODS_W - 0.06$. Despite these models' simplicity, they still explained 64% of total variance in deviation from matching for mice and 45% for monkeys (Monkeys: *Adjusted R*² = 0.45; Mice: *Adjusted R*² = 0.64).

Entropy-based metrics capture the relationship between undermatching and reward environment better than existing metrics. Our previous observation that entropy-based metrics can explain most variance in undermatching behavior suggests that entropy-based metrics may also capture average differences in undermatching between reward schedules in mice and the lack of differences in undermatching between reward schedules in monkeys.

To test whether this was the case, we used 10-fold cross validated linear regression to predict deviation from matching using a model without entropy-based or repetition metrics, a model without repetition metrics, and a model with all metrics. Predictors chosen for inclusion in these models were the predictors that remained in final stepwise regression equations described above.

For mice, in all three models, predicted deviation from matching was significantly lower in the 40/5 than the 40/10 reward environment (**Fig. S5**; two-sided t-test; Model without entropy/repetition metrics: $p = 6.30 \times 10^{-3}$, Model without entropy-based metrics: $p = 3.22 \times 10^{-3}$, Full model with entropy and repetition metrics: $p = 4.76 \times 10^{-25}$). Importantly, the difference between deviation from matching in the two reward schedules was greatest for the full model (**Fig. S5**; Cohen's d ; Model without entropy/repetition metrics: $d = -0.10$, Model without entropy-based metrics: $d = -0.10$, Full model with entropy and repetition metrics: $d = -0.39$). The full model with entropy-based metrics was the only model that came close to replicating the magnitude of differences in deviation from matching between the 40/5 and 40/10 schedules in behavioral data (**Fig. S5c-d**).

For monkeys, predicted deviation from matching from both regression models without entropy metrics was significantly lower in the 70/30 than the 80/20 reward environment (**Fig. S5e-f**; two-sided t-test; Model without entropy/repetition metrics: $p = 1.19 \times 10^{-49}$, Model without entropy-based metrics: $p = 1.32 \times 10^{-29}$). Only the regression model with entropy metrics replicated the observed lack of difference in undermatching between reward schedules. (**Fig. S5g-h**; Full model with entropy metrics: $p = 0.36$; Observed difference between reward schedules: $p = 0.19$). Therefore, entropy-based metrics are necessary and sufficient to capture the influence of reward schedule on deviation from matching.

Reinforcement learning models do not capture entropy-based metrics well. To capture the observed variability in entropy-based metrics and underlying learning and choice mechanisms, we next fit choice behavior using three reinforcement learning (RL) models. These models assumed different updating of reward values and one included a term to mimic the influence of previous choices on choice during the current trial (see **Methods** for more details). Out of the three studied models, we found that RL2, in which the estimated reward values for the unchosen side or stimulus decays to zero over time, provided the best fit of choice behavior for both mice and monkeys as reflected in the lowest Akaike Information Criterion (AIC) (**Fig. 5a-b**).

We next tested whether RL2 could replicate observed distributions of entropy-based metrics and undermatching by simulating the model during our experiment using parameters obtained from model fitting. Due to the large number of simulations performed (100 simulations per session, mice: $n = 331,900$ blocks; monkeys: $n = 221,200$ blocks), we were able to estimate population

distributions of metrics for each model. We found that the average predicted $ERODS_{w-}$ was significantly higher than the average observed $ERODS_{w-}$, suggesting the RL model underutilizes loss-dependent and option-dependent strategies when compared to mice and monkeys in our experiments (**Fig. 5c–d**). Additionally, block-wise predicted and observed $ERODS_{w-}$ were both highly variable, but the distribution of predicted $ERODS_{w-}$ was different than the observed distribution for both mice and monkeys (**Fig. 5e–f**; Two-sided Kolmogorov-Smirnov test; mice: $D = 0.121$; monkeys: $D = 0.072$). Moreover, the predicted distribution of deviation from matching was different from the observed distribution for both mice and monkeys (**Fig 5g–h**; Two-sided Kolmogorov-Smirnov test; mice: $D = 0.091$; monkeys: $D = 0.101$).

Finally, we also computed undermatching and all behavioral metrics in simulated data using RL2 with random parameter values. We found that our entropy-based metrics are better predictors for deviation from matching than the parameters of the RL2 model (see **Fig. S6**). Together, our results illustrate that RL models failed to replicate observed distribution of $ERODS_{w-}$ and variability in matching behavior, pointing to additional mechanisms that contribute to behavior. Moreover, because $ERODS_{w-}$ is highly correlated with undermatching in observed behavior (**Fig. 4**) and RL simulations (**Fig. S6**), our results suggest that a model that better captures $ERODS_{w-}$ may also better capture variability in matching behavior.

Model inspired by entropy-based metrics captures choice behavior more accurately. The deviations of predicted $ERODS_{w-}$ from observed $ERODS_{w-}$ suggest RL models underutilize loss-dependent and option-dependent strategies; that is, they fail to capture the influence of option (or action) and loss in the current trial on choice in the subsequent trial. To fix this problem, we developed a new model, which we refer to as the dynamic loss-choice memory (LCM) model, that augments a RL component with an option-dependent choice memory and an outcome-dependent loss memory (see **Methods**).

Figure 5. Dynamic loss-choice memory model better accounts for choice behavior and undermatching and exhibits less bias and error in capturing entropy-based metrics. (a–b) Comparison of goodness-of-fit of three RL models, the dynamic loss-choice memory (LCM) model, and models with some components of the dynamic LCM model. Plotted is the average Akaike Information Criterion (AIC) over all sessions for mice (a) and monkeys (b). **(c–d)** Comparison of average ERODS_{W-} observed in animals and ERODS_{W-} predicted based on simulations of the best-fitting RL model and the new dynamic LCM model. Error bars represent SEM. **(e–h)** Empirical cumulative distribution functions (CDFs) of ERODS_{W-} (e–f) and deviation from matching (g–h) observed in animals and predicted from simulations of the RL2 dynamic LCM models. Shaded bars around CDFs indicate 95% confidence interval. Insets display the distribution of observed metrics versus metrics predicted using the RL model (left inset) and dynamic LCM model (right inset). Distributions are scaled such that their maximum widths are equal. Displayed *D*-values are the test-statistic from a two-sided Kolmogorov-Smirnov test comparing the distributions. Lower *D*-values indicate increased similarity between distributions (less different CDFs); *D*-values are directly comparable because the same, large number of simulations (100 per session) were performed to estimate population distributions of metrics. The dynamic LCM model better captured deviation from matching in mice by nearly 20% and in monkeys by over 30%.

More specifically, the loss-memory component encourages either staying or switching in response to loss and is modulated by expected uncertainty. Here, expected uncertainty is defined as the expected unsigned reward prediction error in the RL component. As uncertainty increases, the weight of the loss-memory component increases proportionately, making choice more dependent on previous reward feedback. In contrast to RL models with adaptive learning rates (e.g., as in Pearce-Hall model), changes in the loss-memory component only influence subsequent choice, not the aggregation of values in the RL component of our model. The choice-memory component encourages either staying on or switching from options that have been chosen recently. Since the model typically chooses the option with a higher value, the choice-memory component captures strategy in response to selection of the better or worse option. Importantly, our model is agnostic about the strategy employed by the loss- and choice-memory components; for example, the loss-memory component could encourage lose-stay or lose-switch behavior, and the choice-memory component could encourage stay or switch after selection of the worse option (or action), which we refer to as worse-stay and worse-switch. In summary, the loss-memory component can capture lose-switch strategies and the choice-memory component can capture worse-switch strategies. The addition of these components to an RL component creates a model that can flexibly modulate reward- and option-dependent strategies, which may increase the consistency of strategy in response to loss on the worse option (reduce $ERODS_w$).

We found that the dynamic loss-choice memory model fit choice behavior better than all previous RL models as indicated by a lower AIC (**Fig. 5a–b**). Also, the dynamic LCM model explained a large absolute proportion of variance in choice behavior as indicated by McFadden pseudo R^2 values of 0.35 for mice and 0.72 for monkeys (**Table S2**). As expected, the dynamic loss-choice memory model better captured the average consistency of strategy in response to loss on the worse option or action (**Fig. 5c–d**). Importantly, the new model also better captured the observed distribution of $ERODS_w$ in both mice and monkeys (**Fig 5e–f**; two-sided Kolmogorov-Smirnov test; mice: $D = 0.056$; monkeys: $D = 0.040$). This improvement in capturing $ERODS_w$ corresponded with similar improvements in capturing deviation from matching. The predicted distribution of deviation from matching from the dynamic LCM model better replicated the observed distribution of deviation from matching than the predicted distribution from the best-fitting RL model in both mice and monkeys (**Fig 5g–h**; Two-sided Kolmogorov-Smirnov test; mice: $D = 0.072$; monkeys: $D = 0.067$). This improvement is significant; there is

a nearly 20% reduction in the maximum difference between cumulative distribution functions (CDFs) for mice and an over 30% reduction in the maximum difference between CDFs for monkeys.

To identify the contributions of each component of the dynamic LCM model in capturing choice behavior, we also tested four additional models that are composed of RL2 and either the loss-memory or the choice-memory components of the dynamic LCM model. In two of these models, RL2+LM+ and RL2+CM+, the weights of the loss-memory and choice-memory components were restricted to positive values. We found that in mice, the dynamic LCM model fit trial-by-trial choice behavior and captured global undermatching better than models that include either of the two components (**Fig. 5a; Table S2**). Moreover, improvement in capturing local and global choice behaviors in mice was larger when including loss-memory component than choice-memory component, pointing to a larger contribution of loss-dependent strategies in mice. In monkeys, however, goodness-of-fit using the RL2+CM model was indistinguishable from goodness-of-fit based on the dynamic LCM model, suggesting the importance of option-dependent strategy for explaining choice behavior in monkeys (**Fig. 5a; Table S2**). The fit of choice behavior using choice-memory component with positive values only (RL2+CM+ model), was worse than that of the simple RL2 model, whereas bidirectional choice-memory component improved the fit of both local and global choice behaviors (**Fig 5a; Table S2**). These results indicate that bidirectional weighting inspired by entropy-based metrics was responsible for most of the observed improvements in capturing choice behavior in monkeys.

Finally, to understand how the dynamic LCM model modulates specific loss- and option-dependent strategies in mice and monkeys, we compared the distributions of model parameters between the dynamic LCM, RL2, RL2+CM, and RL2+LM models. Interestingly, the loss-memory and choice-memory components both had positive average weights in mice and negative average weights in monkeys (**Fig S7i-l**). In other words, the loss-memory component generally resulted in lose-switch strategies in mice and lose-stay strategies in monkeys, whereas the choice-memory component encouraged worse-switch strategies in mice and worse-stay strategies in monkeys. We also found that in mice, the average weights of the choice- and reward-memory components were higher in the dynamic LCM model than in the RL2+CM and RL2+LM models, respectively, indicating that the two components may interact to modulate

behavior (**Fig S7i–j**). This dovetails with our entropy-based results that show response to the loss after selection of the worse option is uniquely important for global choice behaviors.

To summarize, the proposed model improved fit of choice behavior and captured our metrics and undermatching more accurately, whereby revealing that undermatching behavior arises from competition among multiple components incorporating loss-memory and choice-memory on a trial-by-trial basis. Importantly, we used deviations in predicted entropy-based metrics from their observed values to identify shortcomings in existing RL models and propose mechanisms to mitigate them.

Discussion

Undermatching is a universal behavioral phenomenon that has been observed across many species. Here, we show that novel entropy-based metrics based on response to reward feedback can accurately predict undermatching in mice and monkeys, suggesting that inconsistent use of local reward-dependent and option-dependent strategies can account for a large proportion of variance in global undermatching. Moreover, we demonstrate that these entropy-based metrics can be utilized to construct more complex reinforcement learning models that are able to capture choice behavior, undermatching, and utilization of reward-dependent strategy. Together, our entropy-based metrics provide a powerful model-free tool to develop and refine computational models of choice behavior and reveal neural mechanisms underlying adaptive behavior.

Similar to many previous studies of matching behavior (Herrnstein and Loveland, 1974; William, 1979; Belke and Belliveau, 2001; Sugrue et al., 2004; Tsutsui et al., 2016), we observed significant, but highly variable undermatching in both mice and monkeys. Although many studies (William, 1979; Otto et al., 2011; Iigaya et al., 2019) have speculated a relationship between local choice and learning strategies and global matching behavior, few aimed to predict observed undermatching (except Iigaya et al., 2019). By focusing on the variability in undermatching, here, we were able to show that global undermatching can be largely explained by the degree of inconsistency in response to no reward on the worse option (ERODSw.) across species. Specifically, ERODSw. could explain about 50% and 40% of the variance in undermatching in mice and monkeys, respectively. The proposed entropy metrics were able to predict undermatching across two very different species despite differences in the tasks including

utilized reward probabilities and schedules (40/5 and 40/10 probabilities with baiting vs. complementary 80/20 and 70/30 with no baiting), learning modality (action-based vs. stimulus-based), choice readout (licks vs. saccades), and predictability of block switches (unpredictable vs. semipredictable), suggesting the proposed metrics are powerful and generalizable.

The proposed entropy-based metrics complement and improve upon commonly used behavioral metrics such as win-stay, lose-switch, and the U-measure (Miller and Frick, 1949). Although the U-value has been used to measure consistency or variability in choice behavior (Machado, 1992, 1993), it is difficult to interpret and fails to capture sequential dependencies in choice (Kong et al., 2017). Proposed entropy-based metrics avoid these issues because they have both clear interpretations and can capture the sequential dependence of choice on previous reward and/or selected action or option. Similarly, although win-stay and lose-switch provide valuable information (Bari et al., 2010; Otto et al., 2011; Dalton et al., 2014, 2016; Jang et al., 2015; Gruber and Thapa, 2016), these probabilities do not solely reflect the effects of reward feedback on staying (or similarly switching) as they both depend on the probability of stay. For example, if staying behavior is independent of reward, win-stay and lose-switch values simply reflect the overall stay and switch probabilities, respectively. We also found that win-stay and lose-switch are not strong predictors of undermatching because their relative importance depends on the overall probability of winning. For example, if the overall probability of reward is high, lose-switch is less useful for predicting behavior because response to loss represents strategy in a small subset of trials. Therefore, win-stay and lose-switch cannot capture the degree to which staying and switching strategy depend on reward outcome only. The entropy-based metrics such as ERDS overcome these issues by combining win-stay and lose-switch with $p(\text{win})$ and $p(\text{stay})$. Finally, more complex model-independent metrics involving conditional probabilities (Ito and Doya, 2009) are not desirable because the number of possible conditional probabilities grows exponentially with the number of conditioning variables (e.g., reward, side, stimulus) and the number of preceding trials examined. Entropy-based metrics can overcome this issue by capturing sequential information in a single quantity because additional entropy-based metrics can easily be defined based on longer sequences of choice, reward, other task variables.

As shown by multiple studies, models that fit choice data best may still fail to replicate important aspects of behavior (Palminteri et al., 2017; Wilson and Collins, 2019). Therefore, model

validation must involve analyzing both a model's predictive potential (fitting) and its generative power (replication of behavior in simulations). We used shortcomings of existing RL models in capturing the most predictive entropy-based metrics to detect additional mechanisms underlying adaptive behavior. This approach can be applied to other tasks in which similar or different entropy-based metrics are most predictive of global choice behavior (matching or other metrics). Our aim here was not to find the best model for capturing all aspects of behavior but instead to provide a framework for how response to reinforcement can be used both to explain global behavior and to capture local choice behavior.

Using this method, we constructed a new dynamic loss-choice memory model that augments a reinforcement-learning system with a choice-memory component that captures option-dependent strategies and a loss-memory component that captures loss-dependent strategies. Previous studies have also shown that a combination of win-stay lose-switch (WSLS) strategies with RL models could improve fit of choice behavior and the average matching behavior (Ito and Doya, 2009; Otto et al., 2011; Worthy and Maddox, 2014). The choice-memory component used here is similar to choice kernel models that have been shown to improve fit of choice behavior (Lau and Glimcher, 2005; Wittmann et al., 2020). Nonetheless, the proposed dynamic loss-choice memory model is a novel combination of these components. Critically, the weights of the loss and choice components could be either positive or negative. This parallels how entropy-based metrics capture response to reward feedback considering that low entropy can result from strong positive or negative influences of recent rewards or choices (e.g., high win-stay and high win-switch both correspond to low entropy). Because of this mechanism, the model can explain a few findings not related to our study. For instance, the model can explain how variability in choice can be enhanced (Machado, 1992) by assigning a negative weight to the choice-memory system that encourages switching from recently selected options. Moreover, the model can explain why high levels of global reward can decrease lose-switch (Wittmann et al., 2020): a negative weight in the reward-memory component based on expected unsigned reward prediction error results in increased influence of the component in unrewarded trials (decreasing lose-switch) when global reward rate is high. The model may also facilitate quick adaptation to reversals, a behavior that has previously been explained using Bayesian approaches (Costa et al., 2015), because negative weights in the choice-memory component and in the reward-memory component weighted by unsigned reward prediction error both encourage faster response to reversals. Neural correlates

of a similar loss-memory component weighted by recent reward prediction errors have been identified in the dorsal anterior cingulate cortex of humans (Wittmann et al., 2016). Moreover, neural correlates of such choice memories have been identified in various cortical areas of monkeys including the dorsolateral prefrontal cortex, dorsal medial prefrontal cortex, lateral intraparietal area, and the anterior cingulate cortex (Barraclough et al., 2004; Seo et al., 2009; Spitmaan et al., 2020).

In summary, we show that entropy-based metrics are good predictors of global choice behavior across species and can be used to refine reinforcement learning models. Results from fitting and simulating the dynamic loss-choice memory model suggest that recent choices and rewards affect decisions in ways beyond their influence on the update of the stimulus or action values. Thus, entropy-based metrics have the potential to open a new realm of possibilities for understanding neural mechanisms underlying adaptive behavior.

Methods

Experimental paradigm in mice.

Mice performed a dynamic foraging task in which after receiving a “go cue” signaled by an odor, they licked one of the two water tubes (on left and right) to harvest possible reward. In 5% of trials a “no-go cue” was presented by another odor signaling that licks would not be rewarded or punished. If a mouse licked one of the tubes after a go-cue odor, reward was delivered probabilistically. Each trial was followed by an inter-trial interval drawn from an exponential distribution with a rate parameter of 0.3.

The reward probabilities assigned to the left and right tubes were constant for a fixed number of trials (blocks) and changed throughout the session (block switches). Block lengths were drawn from a uniform distribution that spanned a range of 40–100 trials, however, the exact block lengths spanned smaller ranges for individual sessions, resulting in variable block lengths with most block lengths ranging between 40–80 trials. If mice exhibited strong side-specific biases, block lengths were occasionally shortened or lengthened. “Miss trials,” where the mouse did not make a choice, were excluded for all analyses described here.

Mice performed two versions of the task, one with 16 different sets of reward schedules and another with two sets of reward schedules. The vast majority (469 out of 528) of sessions used two sets of reward probabilities equal to 0.4 and 0.1, and 0.4 and 0.05, which we refer to as 40/10 and 40/5 reward schedules. Here, we focus on the most frequent blocks (40/5 and 40/10 reward schedules). Rewards were “baited” such that if reward was assigned on a given side and that side was not selected, reward would remain on that side until the next time that side was selected. Due to this baiting mechanism, the probability of obtaining reward on the unchosen side increased over time as during foraging in a natural environment. Therefore, optimal performance in the task generally required a strategy that involved selecting options in proportion to their reinforcement rates (perfect “matching” behavior). In total, 16 mice performed 469 sessions of the two-probability version of the task for a total of 3319 blocks (1786 and 1533 blocks with 40/5 and 40/10 reward schedules, respectively) and 189,199 trials. More details about the experimental set up are provided in Bari et al (Bari et al., 2019).

Experimental paradigm in monkeys.

In the reversal learning task in monkeys, Costa and colleagues (Costa et al., 2016) trained monkeys to fixate on a central point on a screen to initiate each trial (**Fig. 1**). After fixation, two stimuli, a square and circle, were presented on the screen to the left and right of the fixation point. The side that the stimuli were presented on was chosen randomly and was not related to reward. Monkeys made saccades to a stimulus and fixated on the stimulus to indicate their choice in each trial. A juice reward was delivered probabilistically via a pressurized tube based on the chosen stimulus. Each trial was followed by a fixed 1.5s inter-trial interval. Trials in which the monkey did not make a choice or failed to fixate were immediately repeated.

Monkeys completed sessions that contained around 1300 trials on average divided into superblocks of 80-trials. Within each superblock the reward probabilities assigned to each cue were reversed randomly between trials 30 and 50, such that the stimuli that was less rewarding at the beginning of the superblock became more rewarding and vice versa. Every 80 trials, monkeys were presented with new stimuli that varied in color but not shape. Monkeys performed two variants of the task, a stochastic variant with three reward schedules (80/20, 70/30, and 60/40) and a deterministic variant with one reward schedule (100/0). Here, we focus our analyses on the 80/20 and 70/30 reward schedules (2212 blocks of the task performed by 4 monkeys) as

they provide two levels of uncertainty similar to the experiment in mice. For more details, please see Costa et al. (Costa et al., 2015, 2016).

Behavioral metrics.

Matching Performance. To measure the overall response to reinforcement on the two choice options (e.g., left and right actions when reward is based on the location) in each block of the experiment, we defined undermatching (UM) as:

$$UM = (Choice_F - Reward_F) \times \text{sign}(Reward_F - 0.5) \quad (\text{Eq. 1})$$

where $\text{sign}(0) = 1$ and choice and reward fractions ($Choice_F$, $Reward_F$) are defined as follows:

$$Choice_F = \frac{P(\text{choosing left})}{P(\text{choosing left}) + P(\text{choosing right})}$$

$$Reward_F = \frac{P(\text{reward on left})}{P(\text{reward on left}) + P(\text{reward on right})} \quad (\text{Eq. 2})$$

Therefore, UM measures the difference between choice and reward fractions toward the more rewarding options (side or stimulus). Similarly, UM can be computed based on the color of stimuli when color is informative about reward outcome. Based on our definition, negative and positive values for UM correspond to undermatching and overmatching, respectively.

Win-stay and lose-switch. Win-stay (WS) and lose-switch (LS) measure the tendency to repeat a rewarded choice (in terms of action or stimulus) and switch away from an unrewarded choice, respectively. These quantities are based on the conditional probabilities of stay and switch after reward and no reward, respectively, and can be calculated in a block of trials as follows:

$$WS = P(\text{stay}|\text{win}) = \frac{P(\text{stay,win})}{P(\text{win})}$$

$$LS = P(\text{switch}|\text{lose}) = \frac{P(\text{switch,lose})}{P(\text{lose})} \quad (\text{Eq. 3})$$

where $P(\text{win})$ and $P(\text{stay})$ are the probabilities of harvesting reward and choosing the same side (or stimulus) in successive trials, $P(\text{lose}) = 1 - P(\text{win})$, and $P(\text{switch}) = 1 - P(\text{stay})$.

Repetition Index. Repetition index (RI) measures the tendency to repeat a choice beyond what is expected by chance and can be computed by subtracting the probability of stay by chance from the original probability of stay (Soltani et al., 2013). RI can be computed based on the repetition of left or right choices as follows:

$$RI_{LR} = P(stay) - (P(left) \times P(left) + P(right) \times P(right)) \quad (\text{Eq. 4})$$

In general, RI reflects a combination of reward-dependent and reward-independent strategies as well as the stochasticity in choice (temperature in the logit function translating value difference to choice probability).

Repetition index can also be measured based on other option or choice attributes that predict reward such as the color of the chosen option. For example, RI can be defined based on selection of the better and worse options (RI_{BW}) when such options exist in a task:

$$\begin{aligned} RI_{BW} &= P(stay) - (P(better) \times P(better) + P(worse) \times P(worse)) \\ &= (P(better(t), better(t-1)) - P(better) \times P(better)) + (P(worse(t), worse(t-1)) - \\ &\quad P(worse) \times P(worse)) \end{aligned} \quad (\text{Eq. 5})$$

where t is the trial number. Using Eq. 5, RI_{BW} can be decomposed into two pieces, RI_B and RI_W , that measure the tendency to repeat the better and worse option, respectively:

$$RI_B = P(better(t), better(t-1)) - P(better) \times P(better) \quad (\text{Eq. 6})$$

$$RI_W = P(worse(t), worse(t-1)) - P(worse) \times P(worse)$$

New entropy-based measures. In order to quantify the influence of previous reward outcome on choice behavior in terms of stay or switch, we defined the conditional entropy of reward-dependent strategies (ERDS) that combine tendencies of win-stay and lose-switch into a single metric. More specifically, ERDS is defined as the conditional entropy of using stay or switch strategy depending on win or lose in the preceding trial:

$$ERDS = H(str|rew) = H(str) - I(str; rew)$$

$$= -\left(P(\text{stay}, \text{win}) \times \log_2\left(\frac{P(\text{stay}, \text{win})}{P(\text{win})}\right) + P(\text{switch}, \text{win}) \times \log_2\left(\frac{P(\text{switch}, \text{win})}{P(\text{win})}\right) + P(\text{stay}, \text{lose}) \times \log_2\left(\frac{P(\text{stay}, \text{lose})}{P(\text{lose})}\right) + P(\text{switch}, \text{lose}) \times \log_2\left(\frac{P(\text{switch}, \text{lose})}{P(\text{lose})}\right)\right) \quad (\text{Eq. 7})$$

where $H(\text{str})$ and $I(\text{str}; \text{rew})$ are entropy of strategy (stay or switch) and mutual information of strategy and reward:

$$\begin{aligned} H(\text{str}) &= -(P(\text{stay}) \times \log_2(P(\text{stay})) + P(\text{switch}) \times \log_2(P(\text{switch}))) \\ I(\text{str}; \text{rew}) &= -\left(P(\text{stay}, \text{win}) \times \log_2\left(\frac{P(\text{stay}, \text{win})}{P(\text{stay}) \times P(\text{win})}\right) + P(\text{switch}, \text{win}) \times \log_2\left(\frac{P(\text{switch}, \text{win})}{P(\text{switch}) \times P(\text{win})}\right) + P(\text{stay}, \text{lose}) \times \log_2\left(\frac{P(\text{stay}, \text{lose})}{P(\text{stay}) \times P(\text{lose})}\right) + P(\text{switch}, \text{lose}) \times \log_2\left(\frac{P(\text{switch}, \text{lose})}{P(\text{switch}) \times P(\text{lose})}\right)\right) \quad (\text{Eq. 8}) \end{aligned}$$

To better show the link between ERDS and win-stay, lose-switch, and $p(\text{win})$, Eq. 7 can be rewritten as follows:

$$ERDS = -(P(\text{win}) \times WS \times \log_2(WS) + P(\text{win}) \times (1 - WS) \times \log_2(1 - WS) + (1 - P(\text{win})) \times (1 - LS) \times \log_2(1 - LS) + (1 - P(\text{win})) \times LS \times \log_2(LS)) \quad (\text{Eq. 9})$$

ERDS can be decomposed into two components, $ERDS_+$ and $ERDS_-$ ($ERDS = ERDS_+ + ERDS_-$), to allow separation of animals' response to rewarded (win) and unrewarded (loss) outcomes:

$$\begin{aligned} ERDS_+ &= H(\text{str}|\text{win}) = -\left(P(\text{stay}, \text{win}) \times \log_2\left(\frac{P(\text{stay}, \text{win})}{P(\text{win})}\right) + P(\text{switch}, \text{win}) \times \log_2\left(\frac{P(\text{switch}, \text{win})}{P(\text{win})}\right)\right) \\ ERDS_- &= H(\text{str}|\text{lose}) = -\left(P(\text{stay}, \text{lose}) \times \log_2\left(\frac{P(\text{stay}, \text{lose})}{P(\text{lose})}\right) + P(\text{switch}, \text{lose}) \times \log_2\left(\frac{P(\text{switch}, \text{lose})}{P(\text{lose})}\right)\right) \quad (\text{Eq. 10}) \end{aligned}$$

The above equations also show that $ERDS_+$ and $ERDS_-$ are linked to win-stay and lose-switch, respectively.

Considering that RI can be decomposed to repetition after better and worse options (Eq. 5), and following the same logic used to derive ERDS, one can define the conditional entropy of option-

dependent strategy (EODS) based on staying on or switching from the better or worse option (or left and right option) in two consecutive trials:

$$\begin{aligned} EODS_{BW} &= H(str|opt) \\ &= -\left(P(stay, better) \times \log_2 \left(\frac{P(stay, better)}{P(better)}\right) + P(switch, better) \times \log_2 \left(\frac{P(switch, better)}{P(better)}\right) + \right. \\ &\quad \left. P(stay, worse) \times \log_2 \left(\frac{P(stay, worse)}{P(worse)}\right) + P(switch, worse) \times \log_2 \left(\frac{P(switch, worse)}{P(worse)}\right)\right) \quad (\text{Eq. 11}) \end{aligned}$$

$EODS_{BW}$ can be decomposed into two components based on the better and worse options:

$$\begin{aligned} EODS_B &= H(str|better) = -\left(P(stay, better) \times \log_2 \left(\frac{P(stay, better)}{P(better)}\right) + \right. \\ &\quad \left. P(switch, better) \times \log_2 \left(\frac{P(switch, better)}{P(better)}\right)\right) \\ EODS_W &= H(str|worse) = -\left(P(stay, worse) \times \log_2 \left(\frac{P(stay, worse)}{P(worse)}\right) + \right. \\ &\quad \left. P(switch, worse) \times \log_2 \left(\frac{P(switch, worse)}{P(worse)}\right)\right) \quad (\text{Eq. 12}) \end{aligned}$$

EODS can also be computed based on other choice attributes (e.g., location or color of options) to examine what information drives stay or switch behavior:

$$\begin{aligned} EODS_{LR} &= H(str|opt) \\ &= -\left(P(stay, right) \times \log_2 \left(\frac{P(stay, right)}{P(right)}\right) + P(switch, right) \times \log_2 \left(\frac{P(switch, right)}{P(right)}\right) + \right. \\ &\quad \left. P(stay, left) \times \log_2 \left(\frac{P(stay, left)}{P(left)}\right) + P(switch, left) \times \log_2 \left(\frac{P(switch, left)}{P(left)}\right)\right) \quad (\text{Eq. 13}) \end{aligned}$$

We should note that ERDS and EODS are directly comparable and provide insight into the consistency of strategy adopted by an animal. Lower ERDS than EODS suggests that an animal's decisions are more consistently influenced by immediate reward feedback than selection of the better or worse option. A lower EODS than ERDS suggests the opposite. It is worth noting that ERDS decompositions (ERDS+ or ERDS-) cannot be directly compared to EODS decompositions ($EODS_B$ or $EODS_W$) because they encompass different sets of trials; that is, trials where the animal wins may not be trials where the animal chooses the better side and vice versa.

Because conditional entropies can be defined for any two discrete random variables, ERDS and EODS can be generalized to combinations or sequences of combinations of reward and option. Hence, we can define the entropy of reward- and option-dependent strategy (ERODS), a measure of the dependence of strategy on option and reward.

$$\begin{aligned}
 ERODS &= H(str|rew, opt) \\
 &= - \left(P(stay, win, better) \times \log_2 \left(\frac{P(stay, win, better)}{P(win, better)} \right) \right. \\
 &\quad + P(stay, win, worse) \times \log_2 \left(\frac{P(stay, win, worse)}{P(win, worse)} \right) \\
 &\quad + P(switch, win, better) \times \log_2 \left(\frac{P(switch, win, better)}{P(win, better)} \right) \\
 &\quad + P(switch, win, worse) \times \log_2 \left(\frac{P(switch, win, worse)}{P(win, worse)} \right) \\
 &\quad + P(stay, lose, better) \times \log_2 \left(\frac{P(stay, lose, better)}{P(lose, better)} \right) \\
 &\quad + P(stay, lose, worse) \times \log_2 \left(\frac{P(stay, lose, worse)}{P(lose, worse)} \right) \\
 &\quad + P(switch, lose, better) \times \log_2 \left(\frac{P(switch, lose, better)}{P(lose, better)} \right) \\
 &\quad \left. + P(switch, lose, worse) \times \log_2 \left(\frac{P(switch, lose, worse)}{P(lose, worse)} \right) \right) \quad (\text{Eq. 14})
 \end{aligned}$$

ERODS can be decomposed based on choosing the better or worse option in the previous trial, winning or losing in the previous trial, or combinations of option and reward (e.g., choose better option and win on the previous trial).

Decomposing ERODS based on reward option combinations gives:

$$\begin{aligned}
 ERODS_{win, better} &= H(str|win, better) \\
 &= - \left(P(stay, win, better) \times \log_2 \left(\frac{P(stay, win, better)}{P(win, better)} \right) \right. \\
 &\quad \left. + P(switch, win, better) \times \log_2 \left(\frac{P(switch, win, better)}{P(win, better)} \right) \right)
 \end{aligned}$$

$$\begin{aligned}
 ERODS_{win,worse} &= H(str|win, worse) \\
 &= - \left(P(stay, win, worse) \times \log_2 \left(\frac{P(stay, win, worse)}{P(win, worse)} \right) \right. \\
 &\quad \left. + P(switch, win, worse) \times \log_2 \left(\frac{P(switch, win, worse)}{P(win, worse)} \right) \right) \\
 ERODS_{lose,better} &= H(str|lose, better) \\
 &= - \left(P(stay, lose, better) \times \log_2 \left(\frac{P(stay, lose, better)}{P(lose, better)} \right) \right. \\
 &\quad \left. + P(switch, lose, better) \times \log_2 \left(\frac{P(switch, lose, better)}{P(lose, better)} \right) \right) \\
 ERODS_{lose,worse} &= H(str|lose, worse) \\
 &= - \left(P(stay, lose, worse) \times \log_2 \left(\frac{P(stay, lose, worse)}{P(lose, worse)} \right) \right. \\
 &\quad \left. + P(switch, lose, worse) \times \log_2 \left(\frac{P(switch, lose, worse)}{P(lose, worse)} \right) \right) \quad (Eq. 15)
 \end{aligned}$$

Finally, ERODS can also be decomposed based on selection of the better or worse option:

$$\begin{aligned}
 ERODS_B &= H(str|rew, better) = ERODS_{win,better} + ERODS_{lose,better} \\
 ERODS_W &= H(str|rew, worse) = ERODS_{win,worse} + ERODS_{lose,worse} \quad (Eq. 16)
 \end{aligned}$$

or winning or losing in the previous trial:

$$\begin{aligned}
 ERODS_+ &= H(str|win, opt) = ERODS_{win,better} + ERODS_{win,worse} \\
 ERODS_- &= H(str|lose, opt) = ERODS_{lose,better} + ERODS_{lose,worse} \quad (Eq. 17).
 \end{aligned}$$

Reinforcement learning models.

We used three reinforcement learning models to fit choice behavior. In all these models, reward probabilities or values associated with the right and left sides (V_{Right}^{RL} and V_{Left}^{RL}) for mice or circle and square stimuli (V_{Circle}^{RL} and V_{Square}^{RL}) for monkeys were updated differently depending on whether a given choice was rewarded or not. In the first model, which we refer to as RL1, only the estimated reward probability associated with the chosen side or stimulus ($V_C^{RL1}(t)$) was updated as follows:

$$V_C^{RL1}(t+1) = V_C^{RL1}(t) + \zeta_C(R_C(t) - V_C^{RL1}(t)) \quad (\text{Eq. 18})$$

where $C \in \{Left, Right\}$ for mice and $C \in \{Circle, Square\}$ for monkeys, $R_C(t) = 1$ or 0 indicates reward outcome on trial t , and ζ_C corresponds to the learning rate (α_{rew} and α_{unrew}) depending on the whether the choice was rewarded or not rewarded.

In the second model (RL2), the estimated reward probability associated with the unchosen side or stimulus ($V_U^{RL2}(t)$) was also updated as follows:

$$V_U^{RL2}(t+1) = V_U^{RL2}(t) - decay_{rate}(V_U^{RL2}(t)) \quad (\text{Eq. 19})$$

where $decay_{rate}$ is the decay (or discount) rate of value of the unchosen side or stimulus, corresponding to forgetting reward values over time. The third model (RL3) was similar to the first model in terms of the integration of reward outcomes and updating of estimated reward probabilities but had an additional choice-memory component for capturing the effects of previous choices. More specifically, the choice-memory value for the chosen side or stimulus in trial t , $V_C^{CM}(t)$, and for the unchosen side or stimulus in trial t , $V_U^{CM}(t)$, are updated as follows:

$$\begin{aligned} V_C^{CM}(t+1) &= V_C^{CM}(t) + \gamma(1 - V_C^{CM}(t)) \\ V_U^{CM}(t+1) &= V_U^{CM}(t) - \gamma(V_U^{CM}(t)) \end{aligned} \quad (\text{Eq. 20})$$

where γ represents the decay rate for the choice-memory value. In RL3, the overall reward value for the chosen (unchosen) side/stimulus on trial t , $V_C(t)$ ($V_U(t)$), is a linear sum of the choice-memory value and the estimated reward probability for RL1:

$$V_{C/U}^{RL3}(t) = V_{C/U}^{RL1}(t) + \omega_{CM} V_{C/U}^{CM}(t) \quad (\text{Eq. 21})$$

where ω_{CM} determines the relative contribution of the choice-memory component and $V_{C/U}^{RL1}$ is the estimated reward probability of the chosen/unchosen side or stimulus based on RL1.

For all three models, the probability of selecting the left (or right) side is represented as a sigmoid function of the difference in estimated reward probabilities or overall reward values for the left and right sides. Hence, the estimated probability of choosing the left side for mice (or circle for monkeys) in trial t , $P_{Left(Circle)}(t)$, is:

$$P_{Left(Circle)}(t) = \left(1 + e^{-\frac{(V_{Left(Circle)}^{RL_i}(t) - V_{Right(Square)}^{RL_i}(t))}{\sigma}} \right)^{-1} \quad (\text{Eq. 22})$$

where σ quantifies the stochasticity of choice and RL_i is RL1, RL2, or RL3. Values of $decay_{rate}$, γ , α_{rew} , and α_{unrew} ranged from 0 to 1 for all three models, and values of σ ranged from 0 to 100. These models were selected because previous research has suggested they can replicate matching or undermatching phenomena (Sugrue et al., 2004; Soltani and Wang, 2006).

As we show here, however, these models did not replicate important behavioral phenomena in our data (see **Results**). Therefore, we developed a new model based on entropy-based metrics. The new model augments RL2 with a loss-memory component to capture loss-dependent strategy and a choice-memory component to capture option-dependent strategy. In the choice-memory component, the choice-memory value of the chosen (unchosen) option ($V_{C/U}^{CM}$) was updated as in RL3, with $\gamma = \text{mean}(\alpha_{rew}, \alpha_{unrew})$ as the decay rate of the choice-memory component.

In the loss-memory component, the value of the chosen side is the expected reward prediction error on the next trial in unrewarded trials and 0 in rewarded trials. Hence, the expected reward prediction error (E_{rpe}) and the loss-memory value of the chosen side ($V_C^{LM}(t)$) are updated as follows:

$$E_{rpe}(t+1) = E_{rpe}(t) + \gamma(|R_C(t) - V_C^{RL2}(t)| - E_{rpe}(t))$$

$$V_C^{LM}(t) = \begin{cases} 0 & \text{if } R_C(t) = 1 \\ -E_{rpe}(t) & \text{if } R_C(t) = 0 \end{cases} \quad (\text{Eq. 24})$$

where $V_C^{RL2}(t)$ is the value of the chosen side in RL2. In the full model (dynamic loss-choice memory, LCM), the overall reward values of the chosen and unchosen sides in trial t , $V_C(t)$ and $V_U(t)$, are computed as follows:

$$V_C(t) = V_C^{RL2}(t) + \omega_{LM} \times V_C^{LM}(t) + \omega_{CM} \times V_C^{CM}(t)$$

$$V_U(t) = V_U^{RL2}(t) + \omega_{CM} \times V_U^{CM}(t) \quad (\text{Eq. 25})$$

where ω_{LM} and ω_{CM} are free parameters that determine the relative weight of the reward-memory and choice-memory component, respectively. Values of ω_{LM} and ω_{CM} varied from -1 to

1, such that the effects of recent reward and choice on future choice could increase either staying or switching behavior. Notably, $V^{RL2}_{C/U}(t)$ are independent of $V^{LM}_C(t)$ and $V^{CM}_C(t)$, such that the values in the reward-memory and choice-memory components only influence choice in the subsequent trial. Finally, the probability of choosing a side for mice (or a stimulus for monkeys) is computed using a sigmoid function of difference in the overall reward values as described above.

To test the components of the full model individually and the effects of allowing negative parameter weights, we fitted four additional control models. The RL2+LM model is the full model without the choice-memory component, and the RL2+CM model is the full model without the loss-memory component. In the RL2+LM+ and RL2+CM+ models, however, ω_{LM} and ω_{CM} range from 0 to 1 instead of from -1 to 1.

Model fitting and simulations.

We used the standard maximum likelihood estimation method to fit and estimate the best-fit parameters for the models described above. To quantify goodness-of-fit, we computed the Akaike Information Criterion (AIC):

$$AIC = -2 \times \log\text{-likelihood} + 2p \quad (\text{Eq. 26})$$

where p is the number of free parameters in a given model.

To quantify an absolute measure of goodness of fit, we computed the McFadden R^2 (McFadden, 1973) for each model:

$$McFadden R^2 = 1 - \frac{\sum_{sessions} \log\text{-likelihood}}{\sum_{sessions} \log\text{-likelihood}_{Null}} = 1 - \frac{\sum_{sessions} \log\text{-likelihood}}{\sum_{sessions} n \times \ln(0.5)} \quad (\text{Eq. 27})$$

where n is the number of trials in a given session. One hundred model simulations were performed per session using best-fit parameters. The large number of simulations allowed us to estimate the population distributions of all metrics. Finally, we conducted additional simulations of RL2 using random parameter values to examine the relationship between parameters and entropy-based metrics. For these simulations, α_{rew} , α_{unrew} , and σ varied in the range of (0, 1) and $decay_{rate}$ was set to 0.1.

Data analyses and stepwise regressions.

Stepwise regressions were conducted using MATLAB's `stepwiselm` and `stepwisefit` functions. The criterion for adding or removing terms from the model was based on an F-test of the difference in sum of squared error resulting from the addition or removal of a term from the model. A predictor was added to the model if the p -value of the F-test was less than 0.0001, and a predictor was removed from the model if the p -value of the F-test was greater than 0.00011.

The order in which predictors were added to the stepwise regression models was as follows.

Model without entropy/repetition metrics (mice): In the first step of the stepwise regression, $p(win)$ entered the regression equation as a significant predictor of undermatching, $RMSE = 0.1119, p = 4.12 \times 10^{-175}$. In the next steps, $p(stay)$ ($RMSE = 0.1053, p = 2.41 \times 10^{-89}$) and $p(stay|win)$ ($RMSE = 0.1046, p = 7.41 \times 10^{-11}$) entered the regression equation.

Model without entropy/repetition metrics (monkeys): In the first step of the stepwise regression, $p(stay)$ entered the regression equation as a significant predictor of undermatching, $RMSE = 0.0599, p = 1.31 \times 10^{-160}$. In the next steps, $p(win)$ ($RMSE = 0.0579, p = 1.86 \times 10^{-35}$), $p(switch|lose)$ ($RMSE = 0.0575, p = 2.85 \times 10^{-8}$), and $p(stay|win)$ ($RMSE = 0.0571, p = 5.61 \times 10^{-9}$) entered the regression equation. In the final step, $p(stay)$ was removed from the equation because it was no longer a significant predictor ($RMSE = 0.0572, p = 6.37 \times 10^{-4}$).

Model without entropy metrics (mice): In the first step of the regression process, $p(win)$ entered the model ($RMSE = 0.1119, p < 4.12 \times 10^{-175}$). In the next steps, $p(stay)$ ($RMSE = 0.1053, p < 2.41 \times 10^{-89}$), RI_W ($RMSE = 0.0929, p < 4.80 \times 10^{-181}$), $p(stay|win)$ ($RMSE = 0.0914, p < 1.03 \times 10^{-24}$), and RI_B ($RMSE = 0.0910, p < 2.98 \times 10^{-8}$) were added to the regression equation.

Model without entropy metrics (monkeys): In the first step of the regression process, $p(stay)$ entered the model ($RMSE = 0.0599, p = 1.31 \times 10^{-160}$). In the next steps,

$RI_B(RMSE = 0.0505, p = 6.48 \times 10^{-167})$ and $p(\text{switch}|\text{lose})(RMSE = 0.0503, p < 8.59 \times 10^{-5})$ were added to the regression equation.

Full model (mice): In the first step of the regression process, $ERODS_{W-}$ entered the regression equation as a significant predictor of deviation from matching, $RMSE = 0.0717, p < 10^{-300}$. Next, $ERODS_{W+}$ entered the regression equation, $RMSE = 0.0681, p < 1.61 \times 10^{-65}$. In the following steps, $ERODS_{B+}$ ($RMSE = 0.0666, p < 1.97 \times 10^{-29}$), $ERDS_{+}$ ($RMSE = 0.0658, p < 4.11 \times 10^{-18}$), $EODS_W$ ($RMSE = 0.0651, p < 1.72 \times 10^{-15}$), $P(\text{switch}|\text{lose})$ ($RMSE = 0.0639, p < 7.54 \times 10^{-24}$), RI_W ($RMSE = 0.0617, p < 1.46 \times 10^{-46}$), $P(\text{stay}|\text{win})$ ($RMSE = 0.0585, p < 1.81 \times 10^{-67}$), $ERODS_{B-}$ ($RMSE = 0.0573, p < 5.15 \times 10^{-28}$), $ERDS_{-}$ ($RMSE = 0.0568, p < 9.59 \times 10^{-13}$), $p(\text{win})$ ($RMSE = 0.0566, p < 3.75 \times 10^{-6}$), and $EODS_B$ ($RMSE = 0.0564, p < 3.28 \times 10^{-5}$) were added to the regression equation. In the final step, $ERODS_{B+}$ ($RMSE = 0.0565, p < 5.32 \times 10^{-3}$) was removed from the equation.

Full model (monkeys): In the first step of the regression process, $ERODS_{W-}$ entered the regression equation as a significant predictor of deviation from matching, $RMSE = 0.0589, p < 7.31 \times 10^{-114}$. Next, $EODS_W$ entered the regression equation, $RMSE = 0.0551, p < 1.50 \times 10^{-41}$. In the following steps, $ERODS_{B+}$ ($RMSE = 0.0525, p < 2.15 \times 10^{-31}$), RI_B ($RMSE = 0.0512, p < 1.09 \times 10^{-16}$), $P(\text{switch}|\text{lose})$ ($RMSE = 0.0477, p < 1.10 \times 10^{-43}$), $P(\text{stay}|\text{win})$ ($RMSE = 0.0468, p < 2.02 \times 10^{-13}$), and $ERDS_{+}$ ($RMSE = 0.0463, p < 5.44 \times 10^{-8}$) were added to the regression equation.

We note that there were fewer blocks used in the full model stepwise regression because some of the specific entropy metrics were not defined for certain blocks, e.g., if a mouse or monkey never won on the worse side in a block, then $ERODS_{W+}$ was undefined for that block. This resulted in the exclusion of around 500 blocks for mice and 700 blocks for monkeys in the final regression.

We also conducted ten-fold cross-validated regressions to predict deviation from matching (**Fig. S5**) using MATLAB's `fitrlinear` and `kfoldPredict` functions. More specifically, stepwise regressions were performed on a set of possible predictors to determine which predictors to include in the final regression model. Then, cross-validated regressions were computed to predict

deviation from matching using the set of predictors included in the final stepwise regression model.

References

- Anderson, K. G., Velkey, A. J. & Woolverton, W. L. The generalized matching law as a predictor of choice between cocaine and food in rhesus monkeys. *Psychopharmacology (Berl.)* **163**, 319–326 (2002).
- Bari, A. *et al.* Serotonin modulates sensitivity to reward and negative feedback in a probabilistic reversal learning task in rats. *Neuropsychopharmacology* **35**, 1290–1301 (2010).
- Bari, B. A. *et al.* Stable Representations of Decision Variables for Flexible Behavior. *Neuron* **103**, 922–933.e7 (2019).
- Barracough, D. J., Conroy, M. L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nat. Neurosci.* **7**, 404–410 (2004).
- Beardsley, S. D. & McDowell, J. J. Application of Herrnstein’s hyperbola to time allocation of naturalistic human behavior maintained by naturalistic social reinforcement. *J. Exp. Anal. Behav.* **57**, 177–185 (1992).
- Belke, T. W. & Belliveau, J. The general matching law describes choice on concurrent variable-interval schedules of wheel-running reinforcement. *J. Exp. Anal. Behav.* **75**, 299–310 (2001).
- Cero, I. & Falligant, J. M. Application of the generalized matching law to chess openings: A gambit analysis. *J. Appl. Behav. Anal.* **53**, 835–845 (2020).
- Costa, V. D., Tran, V. L., Turchi, J., & Averbeck, B. B. Reversal Learning and Dopamine: A Bayesian Perspective. *J. Neurosci.* **35**, 2407–2416 (2015).
- Costa, V. D., Dal Monte, O., Lucas, D. R., Murray, E. A. & Averbeck, B. B. Amygdala and Ventral Striatum Make Distinct Contributions to Reinforcement Learning. *Neuron* **92**, 505–517 (2016).
- Dalton, G. L., Phillips, A. G. & Floresco, S. B. Preferential involvement by nucleus accumbens shell in mediating probabilistic learning and reversal shifts. *J. Neurosci.* **34**, 4618–4626 (2014).
- Dalton, G. L., Wang, N. Y., Phillips, A. G. & Floresco, S. B. Multifaceted Contributions by Different Regions of the Orbitofrontal and Medial Prefrontal Cortex to Probabilistic Reversal Learning. *J. Neurosci.* **36**, 1996–2006 (2016).

- Fonseca, M. S., Murakami, M. & Mainen, Z. F. Activation of Dorsal Raphe Serotonergic Neurons Promotes Waiting but Is Not Reinforcing. *Curr. Biol.* **25**, 306–315 (2015).
- Gallistel, C. R. *et al.* Is Matching Innate? *J. Exp. Anal. Behav.* **87**, 161–199 (2007).
- Gallistel, C. R. Foraging for brain stimulation: toward a neurobiology of computation. *Cognition* **50**, 151–170 (1994).
- Grossman, C. D., Bari, B. A. & Cohen, J. Y. Serotonin neurons modulate learning rate through uncertainty. *bioRxiv* (2020) doi:10.1101/2020.10.24.353508.
- Gruber, A. J. & Thapa, R. The Memory Trace Supporting Lose-Shift Responding Decays Rapidly after Reward Omission and Is Distinct from Other Learning Mechanisms in Rats. *eNeuro* **3**, (2016).
- Herrnstein, R. J. Relative and absolute strength of response as a function of frequency of reinforcement. *J. Exp. Anal. Behav.* **4**, 267–272 (1961).
- Herrnstein, R. J. & Loveland, D. H. Hunger and contrast in a multiple schedule. *J. Exp. Anal. Behav.* **21**, 511–517 (1974).
- Iigaya, K. *et al.* Deviation from the matching law reflects an optimal strategy involving learning over multiple timescales. *Nat. Commun.* **10**, 1466 (2019).
- Iigaya, K. & Fusi, S. Dynamical regimes in neural network models of matching behavior. *Neural Comput.* **25**, 3093–3112 (2013).
- Ito, M. & Doya, K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *J. Neurosci.* **29**, 9861–9874 (2009).
- Jang, A. I. *et al.* The Role of Frontal Cortical and Medial-Temporal Lobe Brain Areas in Learning a Bayesian Prior Belief on Reversals. *J. Neurosci.* **35**, 11751–11760 (2015).
- Kong, X., McEwan, J. S., Bizo, L. A. & Foster, T. M. An Analysis of U-Value as a Measure of Variability. *Psychol. Rec.* **67**, 581–586 (2017).
- Kubaneck, J. & Snyder, L. H. Matching Behavior as a Tradeoff Between Reward Maximization and Demands on Neural Computation. *F1000Research* **4**, 147 (2015).
- Lau, B. & Glimcher, P. W. Dynamic Response-by-Response Models of Matching Behavior in Rhesus Monkeys. *J. Exp. Anal. Behav.* **84**, 555–579 (2005).
- Lee, S.-H. *et al.* Neural Signals Related to Outcome Evaluation Are Stronger in CA1 than CA3. *Front. Neural Circuits* **11**, 40 (2017).
- Loewenstein, Y. & Seung, H. S. Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 15224–15229 (2006).

- Machado, A. Behavioral variability and frequency-dependent selection. *J. Exp. Anal. Behav.* **58**, 241–263 (1992).
- Machado, A. Learning variable and stereotypical sequences of responses: Some data and a new model. *Behav. Processes* **30**, 103–129 (1993).
- Mazur, J. E. Optimization theory fails to predict performance of pigeons in a two-response situation. *Science* **214**, 823–825 (1981).
- McFadden, D. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, 105–142 (Academic Press, 1973).
- Miller, G. A. & Frick, F. C. Statistical behavioristics and sequences of responses. *Psychol. Rev.* **56**, 311–324 (1949).
- Otto, A. R., Taylor, E. G. & Markman, A. B. There are at least two kinds of probability matching: evidence from a secondary task. *Cognition* **118**, 274–279 (2011).
- Palminteri, S., Wyart, V. & Koechlin, E. The Importance of Falsification in Computational Cognitive Modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).
- Pierce, W. D. & Epling, W. F. Choice, matching, and human behavior: A review of the literature. *Behav. Anal.* **6**, 57–76 (1983).
- Saito, H., Katahira, K., Okanoya, K. & Okada, M. Bayesian deterministic decision making: a normative account of the operant matching law and heavy-tailed reward history dependency of choices. *Front. Comput. Neurosci.* **8**, 18 (2014).
- Savastano, H. I. & Fantino, E. Human choice in concurrent ratio-interval schedules of reinforcement. *J. Exp. Anal. Behav.* **61**, 453–463 (1994).
- Schroeder, S. R. & Holland, J. G. Reinforcement of eye movement with concurrent schedules. *J. Exp. Anal. Behav.* **12**, 897–903 (1969).
- Seo, H., Barraclough, D. J. & Lee, D. Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. *J. Neurosci.* **29**, 7278–7289 (2009).
- Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
- Soltani, A., Noudoost, B. & Moore, T. Dissociable dopaminergic control of saccadic target selection and its implications for reward modulation. *Proc. Natl. Acad. Sci.* **110**, 3579–3584 (2013).
- Soltani, A., Rakhshan, M., Schafer, R. J., Burrows, B. E. & Moore, T. Separable Influences of Reward on Visual Processing and Choice. *J. Cogn. Neurosci.* **33**, 248–262 (2021).

- Soltani, A. & Wang, X.-J. A Biophysically Based Neural Model of Matching Law Behavior: Melioration by Stochastic Synapses. *J. Neurosci.* **26**, 3731–3744 (2006).
- Spitmaam, M., Seo, H., Lee, D. & Soltani, A. Multiple timescales of neural dynamics and integration of task-relevant signals across cortex. *Proc. Natl. Acad. Sci.* **117**, 22522–22531 (2020).
- Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Matching behavior and the representation of value in the parietal cortex. *Science* **304**, 1782–1787 (2004).
- Tsutsui, K.-I., Grabenhorst, F., Kobayashi, S. & Schultz, W. A dynamic code for economic object valuation in prefrontal cortex neurons. *Nat. Commun.* **7**, 12554 (2016).
- Villarreal, M. *et al.* Bayesian methods applied to the generalized matching law. *J. Exp. Anal. Behav.* **111**, 252–273 (2019).
- de Villiers, P. A. & Herrnstein, R. J. Toward a law of response strength. *Psychol. Bull.* **83**, 1131–1153 (1976).
- Vullings, C. & Madelain, L. Control of saccadic latency in a dynamic environment: allocation of saccades in time follows the matching law. *J. Neurophysiol.* **119**, 413–421 (2018).
- William, B. M. Matching, Undermatching, and Overmatching in Studies of Choice. *J. Exp. Anal. Behav.* **32**, 269–281 (1979).
- Williams, B. A. Reinforcement, choice, and response strength. in *Stevens' handbook of experimental psychology* vol. 2 167–244 (John Wiley & Sons, 1988).
- Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *eLife* **8**, (2019).
- Wittmann, M. K. *et al.* Global reward state affects learning and activity in raphe nucleus and anterior insula in monkeys. *Nat. Commun.* **11**, 3771 (2020).
- Wittmann, M. K. *et al.* Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex. *Nat. Commun.* **7**, 12327 (2016).
- Worthy, D. A. & Maddox, W. T. A Comparison Model of Reinforcement-Learning and Win-Stay-Lose-Shift Decision-Making Processes: A Tribute to W.K. Estes. *J. Math. Psychol.* **59**, 41–49 (2014).

Figure S1. Undermatching and behavioral metrics depend on reward probabilities. Plotted are the probability of winning (a), the probability of choosing the better option (b), probability of staying on previous choice (c), win-stay (d), lose-switch (e), a measure of undermatching (f), and RI and RI decompositions (g-h) in the 40/10 and 40/5 reward schedules for mice and the 70/30 and 80/20 reward schedules for monkeys. Error bars indicate s.e.m. For mice, $n_{40/5} = 1786$ blocks and $n_{40/10} = 1533$ blocks, and for monkeys, $n_{70/30} = 1110$ blocks and $n_{80/20} = 1102$ blocks. The asterisk indicates a significant difference between the two environments using two-sided t-test with p -value and Cohen's d -value reported on each panel. In mice, the probability of winning was significantly higher in the 40/10 schedule despite a lower probability of choosing the better side. In monkeys, the probability of winning and probability of choosing better was higher in the 80/20 schedule. Moreover, the probability of staying and the repetition index were both significantly lower in the 40/10 schedule in mice and 70/30 schedule in monkeys because the reward probabilities for the two options are more similar. Finally, both win-stay and lose-switch were closer to 0.5 in the 40/10 schedule for mice, and win-stay was closer to 0.5 in the 70/30 schedule for monkeys which may indicate a decrease in the dependence of staying and switching behavior on reward. However, the differences in $p(\text{win})$ and $p(\text{stay})$ between these environments make interpreting win-stay and lose-switch in isolation challenging.

Figure S2. Entropy-based metrics capture changes in learning strategy between the two reward schedules. Plotted are ERDS and ERDS decompositions (a–c), EODS and EODS decompositions (d–f), and ERODS and ERODS decompositions (g–i) in the 40/10 and 40/5 reward schedules for mice and the 70/30 and 80/20 reward schedules for monkeys. Error bars indicate s.e.m. Reported are the p -values (two-sided t-test) and Cohen's d -values. ERDS, EODS, and ERODS were significantly higher in the 40/10 schedule in mice and the 70/30 schedule in monkeys. Overall, increased entropy in the 40/10 schedule in mice and the 70/30 schedule in monkeys suggests a decrease in the consistency of reward and option-dependent strategy. Decreased consistency of reward and option-dependent strategy may be due to the greater similarity of reward probabilities for the two options in the 40/10 and 70/30 reward schedules.

Figure S3. Correlation between undermatching and proposed entropy-based metrics separately for each reward schedule in mice. Correlation matrix for 19 behavioral metrics and undermatching using Pearson (a) and Spearman (b) tests, separately for blocks with reward probabilities equal to 40/10 (a–b) or 40/5 (c–d). Correlation coefficients are computed across all blocks, and matrix elements with non-significant values ($p > .0001$) are not shown (cells in black). The red rectangles highlight correlation coefficients between behavioral metrics and undermatching. Overall, the new entropy-based metrics were similarly strongly correlated with undermatching in both reward environments.

Figure S4. Correlation between undermatching and proposed entropy-based metrics separately for each reward schedule in monkeys. Same as Figure S3 but for monkeys, with blocks with reward probabilities equal to 70/30 (a–b) or 80/10 (c–d). Overall, the new entropy-based metrics were similarly strongly correlated with undermatching in both reward environments.

Figure S5. Entropy-based metrics capture differences in undermatching between reward environments. (a–d) Plotted are deviation from matching values in the 40/10 and 40/5 reward schedules for mice predicted from 10-fold cross-validated linear regression models using all behavioral metrics except entropy and repetition metrics as predictors (a), all behavioral metrics except entropy-based metrics as predictors (b), and all metrics as predictors (c) versus observed deviation from matching (d). Predictors were selected for inclusion using the stepwise regressions described in the manuscript, then 10-fold cross-validated linear regression was performed to fit models and predict undermatching (U.M.) (see **Methods**). Error bars indicate s.e.m. Reported are the p -values (two-sided t-test) on the left and Cohen's d -values on the right. (e–h) Similar to (a–d), but with monkey data in the 70/30 and 80/20 reward schedules. The full regression model replicates the observed differences between reward schedules in mice and the lack of observed differences between reward schedules in monkeys.

Figure S7. Distributions of model parameters obtained from fitting choice behavior. Plotted are distributions of the learning rate on the rewarded option (a,c), stochasticity of choice (b,d), learning rate on the unrewarded action/option (e,g), decay rate (f,h), the weight of the choice memory component (i,j), and the weight of the loss memory component (k,l) for mice (a-j) and monkey (c-l) using four reinforcement learning models as noted in the legend. Distribution curves are estimated using kernel smoothing with reflection for boundary correction. Dashed vertical lines indicate the mean of each distribution in all plots and the black dashed vertical lines in (i-l) are zero lines. In mice, the choice and loss memory components have positive weights, whereas in monkeys the choice and loss memory components have slightly negative weights.

Metric	Decomposition	Description	Components
ERDS		Entropy of reward -dependent strategy.	$p(\text{win})$, $p(\text{stay} \text{win})$, $p(\text{switch} \text{lose})$
ERDS₊		Entropy of win -dependent strategy.	$p(\text{win})$, $p(\text{stay} \text{win})$
ERDS₋		Entropy of loss -dependent strategy	$p(\text{lose})$, $p(\text{switch} \text{lose})$
EODS		Entropy of option -dependent strategy	$p(\text{choose better})$, $p(\text{stay} \text{choose better})$, $p(\text{switch} \text{choose worse})$
EODS_B		Entropy of option -dependent strategy on the better side .	$p(\text{choose better})$, $p(\text{stay} \text{choose better})$
EODS_w		Entropy of option -dependent strategy on the worse side .	$p(\text{choose worse})$, $p(\text{switch} \text{choose worse})$
ERODS		Entropy of reward- and option -dependent strategy	$p(\text{win, better})$, $p(\text{win, worse})$, $p(\text{lose, better})$, $p(\text{lose, worse})$, $p(\text{stay} \text{win, better})$, $p(\text{stay} \text{win, worse})$, $p(\text{switch} \text{lose, better})$, $p(\text{switch} \text{lose, worse})$
ERODS_{B+}		Entropy of reward- and option -dependent strategy in response to win after selecting the better option .	$p(\text{win, better})$, $p(\text{stay} \text{win, better})$
ERODS_{B-}		Entropy of reward- and option -dependent strategy in response to loss after selecting the better option .	$p(\text{lose, better})$, $p(\text{switch} \text{lose, better})$
ERODS_{w+}		Entropy of reward- and option -dependent strategy in response to win after selecting the worse option .	$p(\text{win, worse})$, $p(\text{stay} \text{win, worse})$
ERODS_{w-}		Entropy of reward- and option -dependent strategy in response to loss after selecting the worse option .	$p(\text{lose, worse})$, $p(\text{switch} \text{lose, worse})$

Table S1. Summary and components of entropy-based metrics. Each row contains a short description of a metric, its decompositions, and the set of component probabilities that can be used to compute the metric.

Model	Model Description	Parameters	-LL	MF R^2	AIC	D ERODSw-	D matching
RL1	Return-based RL	$\alpha_{rew}, \alpha_{unrew}, \beta$	228.31	0.184	462.62 (88.97*)	0.238	0.224
			21.05	0.620	48.11 (4.47*)	0.054	0.092
RL2	Income-based RL	$\alpha_{rew}, \alpha_{unrew}, \beta, \text{decay}_{rate}$	189.61	0.322	387.22 (12.57*)	0.121	0.091
			18.09	0.674	44.18 (0.54*)	0.072	0.101
RL3	Return-based RL + full choice kernel	$\alpha_{rew}, \alpha_{unrew}, \beta, \omega_{CM}, \gamma_{CM}$	190.72	0.318	391.44 (16.79*)	0.095	0.088
			18.24	0.671	46.47 (2.83*)	0.027	0.072
RL2+ CM	Income-based RL + choice memory	$\alpha_{rew}, \alpha_{unrew}, \beta, \text{decay}_{rate}, \omega_{CM}$	187.83	0.328	385.65 (11.00*)	0.119	0.090
			16.54	0.702	43.10 (-0.54*)	0.037	0.065
RL2+ CM+	Income-based RL + choice memory with positive CM weight	$\alpha_{rew}, \alpha_{unrew}, \beta, \text{decay}_{rate}, \omega_{CM}$	188.37	0.326	386.74 (12.09*)	0.122	0.090
			17.90	0.677	45.79 (2.15*)	0.072	0.101
RL2+ LM	Income-based RL + loss memory	$\alpha_{rew}, \alpha_{unrew}, \beta, \text{decay}_{rate}, \omega_{LM}$	183.13	0.345	376.26 (1.61*)	0.060	0.079
			17.06	0.692	44.13 (0.49*)	0.060	0.089
RL2+ LM+	Income-based RL + loss memory with positive LM weight	$\alpha_{rew}, \alpha_{unrew}, \beta, \text{decay}_{rate}, \omega_{LM}$	183.16	0.345	376.31 (1.66*)	0.059	0.078
			17.07	0.692	44.14 (0.50*)	0.060	0.089
Dyn. LCM	Income-based RL + loss memory + choice memory	$\alpha_{rew}, \alpha_{unrew}, \beta, \text{decay}_{rate}, \omega_{CM}, \omega_{LM}$	181.32	0.352	374.65	0.056	0.072
			15.82	0.715	43.64	0.040	0.067

Table S2. Various models used to fit choice data, their parameters and goodness-of-fit measures, and models' ability to capture behavioral metrics. Each row provides a short description of a given model, its parameters, goodness-of-fit based on the negative log-likelihood values separately for mice and monkeys (column 4), McFadden R^2 values or variance in choice explained by each model (column 5), goodness-of-fit based on the AIC (column 6), D-values based on Kolmogorov-Smirnov tests comparing distributions of ERODSw- (column 7) and deviation from matching (column 8) predicted using model simulations and actual behavior. Values reported in parentheses and the asterisks in column 6 indicate the difference in AIC of a given model and the full model (Dyn. LCM) and the significance of this difference using paired-sample t-test ($p < 3.0 \times 10^{-5}$). Rows in orange and cyan correspond to mouse and monkey data, respectively.