Article

Not peer-reviewed version

An Information-Theoretic Framework for Understanding Learning and Choice Under Uncertainty

Jae Hyung Woo , Lakshana Balaji , Alireza Soltani *

Posted Date: 2 September 2025

doi: 10.20944/preprints202509.0001.v1

Keywords: reinforcement learning; value-based decision making; conditional entropy; mutual information; uncertainty

Article

An Information-Theoretic Framework for Understanding Learning and Choice Under Uncertainty

Jae Hyung Woo [1], **Lakshana Balaji** [2] **and Alireza Soltani** [1,*]

[1] Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755, USA
[2] Department of Biology, Indian Institute of Science Education and Research Tirupati (IISER T), Panguru, Tirupati, India
[*] Correspondence: alireza.soltani@dartmouth.edu

Abstract

Although information theory is widely used in neuroscience, its application has primarily been limited to the analysis of neural activity, with much less emphasis on behavioral data. This is despite the fact that the discrete nature of behavioral variables in many experimental settings—such as choice and reward outcomes—makes them particularly well-suited to information-theoretic analysis. In this study, we provide a framework for how behavioral metrics based on conditional entropy and mutual information can be used to infer an agent's decision-making and learning strategies under uncertainty. Using simulated reinforcement-learning models as ground truth, we illustrate how information-theoretic metrics can reveal the underlying learning and choice mechanisms. Specifically, we show that these metrics can uncover: (1) a positivity bias, reflected in higher learning rates for rewarded compared to unrewarded outcomes; (2) gradual, history-dependent changes in the learning rates indicative of metaplasticity; (3) adjustments in choice strategies driven by reward harvest rate; and (4) presence of alternative learning strategies and their interaction. Overall, our study highlights how information theory can leverage the discrete, trial-by-trial structure of many cognitive tasks, offering a versatile framework for investigating neural and computational mechanisms of learning and choice under uncertainty—with potential for further extension.

Keywords: reinforcement learning; value-based decision making; conditional entropy; mutual information; uncertainty

1. Introduction

Information theory has been used widely across different domains of cognitive and systems neuroscience, ranging from analyzing spike trains to estimating uncertainty in sensory environments. For example, Shannon entropy and related metrics have been used to study information flow [1,2], functional and effective connectivity [3–7], and variability in neural response [8–11]. These quantitative approaches have provided significant insight into brain functions at the cellular and system levels.

Despite this widespread use of information-theoretic metrics in analyzing neural data, their application in investigating behavior has been surprisingly limited. Although several studies have employed entropy-based metrics to quantify uncertainty in stimuli [12–15] and outcomes [16–18], there are only a few studies that have utilized information theory to directly examine the underlying decision-making and learning mechanisms [19–21]. This is despite the fact that many behavioral measures—such as binary choices and reward feedback—are ideally structured for analysis using information-theoretic tools.

A common tool for studying learning and choice behavior is reinforcement learning (RL) models [22–27], due to their simplicity and interpretability. In this approach, various RL models are constructed and fit to the choice data. The best-fitting model is then identified through model selection, and its components form the basis for inferences about underlying cognitive and/or neural mechanisms.

However, these models often need to be augmented with additional components to capture empirical data in specific tasks (e.g., [28–31]), while it remains unclear how many components are necessary—or sufficient—to define the "best" model. Such extensions include separate learning rates for positive versus negative prediction errors (differential learning rates), time-varying (dynamic) learning rates, modulations of learning by reward harvest rate, and arbitration between alternative models of the environment, among others [32–40]. Nevertheless, currently there is no systematic method to identify the critical components required in RL models. This lack of methodology can lead to important mechanisms being overlooked, even in models that provide the best fit among the tested models.

Interestingly, choice behavior under uncertainty is inherently stochastic, making it well-suited for analysis using information theory. Recently, a set of information-theoretic metrics was proposed as a model-agnostic approach to uncover the hidden learning mechanisms underlying behavior. For example, Trepka et al. [29] have suggested that behavioral metrics based on conditional entropy can quantify the consistency in local choice strategies and predict undermatching. Moreover, Woo et al. [41] have shown that mutual information, alongside other measures, can capture the influence of high-order reward statistics. These results highlight the potential of information-theoretic metrics to probe learning and decision-making processes beyond what traditional model-fitting techniques can reveal.

Here, we extended this approach by applying information-theoretic metrics to simulated choice data from different RL models—serving as ground truth— across four learning tasks. We show that this method captures keys aspects of learning and decision making without relying on model fitting. To that end, we used existing metrics and developed new ones to detect a range of learning and decision-making mechanisms and their dynamic adjustments. We begin by identifying higher learning rates following positive prediction errors (rewarded outcomes) than following negative prediction errors (unrewarded outcomes)—a phenomenon referred to as positivity bias [42–45]. We also examine changes in learning rates over time as a result of metaplasticity [46,47]. Next, we investigate the influence of reward harvest rate, which has been shown to modulate learning and decision making [34,35,48,49]. Finally, in naturalistic reward environments, choice options often possess multiple attributes, each potentially predictive of reward outcomes. Previous studies have demonstrated that humans and other animals tackle this challenge by simultaneously learning about alternative reward contingencies (i.e., models of the environment), arbitrating between these models, and deploying selective attention to guide differential learning and decision making [38,50–58]. Therefore, we test how information-theoretic metrics can be used to detect the presence of distinct learning strategies in complex reward environments. Overall, our results demonstrate that the patterns of information-theoretic metrics provide useful summary statistics of the behavioral signatures generated by different learning and choice mechanisms, thereby offering a complementary approach to model fitting (both discovery and recovery).

2. Materials and Methods

2.1. General Experimental Paradigm: Probabilistic Reversal Learning

We simulated choice data using different reinforcement learning (RL) agents performing two variants of the probabilistic reversal learning task (PRL), a widely used paradigm to assess cognitive flexibility across various species [59]. In a typical probabilistic reversal learning paradigm, subjects choose between two options based on reward feedback they receive on each trial, with selection of each option associated with a different probability of reward. Choice options can take various forms, such as distinct visual stimuli on a computer screen, identical stimuli presented at different spatial locations, or physical levers. One option yields reward with a higher probability (the better option) than the other (the worse option), but these reward contingencies switch at fixed or random times within an experimental session, creating 'reversals'. Crucially, reversals are not signaled and thus, subjects have to adjust their choice solely based on reward feedback to maximize their chance of winning a reward. Commonly used rewards include drops of juice or water (e.g., [60,61]), sucrose pellets (e.g., [62,63]), and monetary incentives for humans (e.g., [56,64]), which are delivered probabilistically at the end of each trial. For most simulations, we set the reward probabilities at 80 and 20 (corresponding to 80/20% reward schedule) with each block consisting of 80 trials, unless stated otherwise. Each block contains a reversal in the middle of the block, where reward probabilities on the two options switch. In the final set of simulations (Section 3.4), we also considered a generalized PRL task in which reward probabilities depended on one feature of each choice option (e.g., its color or its shape), while the other feature carried no information about reward.

2.2. Information-Theoretic Metrics

In this study, we utilized and extended the information-theoretic metrics introduced previously [29,41]. These measures——grounded in conditional entropy, mutual information, and outcome-specific decompositions——quantify how past rewards and choices shape the uncertainty surrounding an agent's decision to stay with or switch from its previous choice (**Figure 1b**).

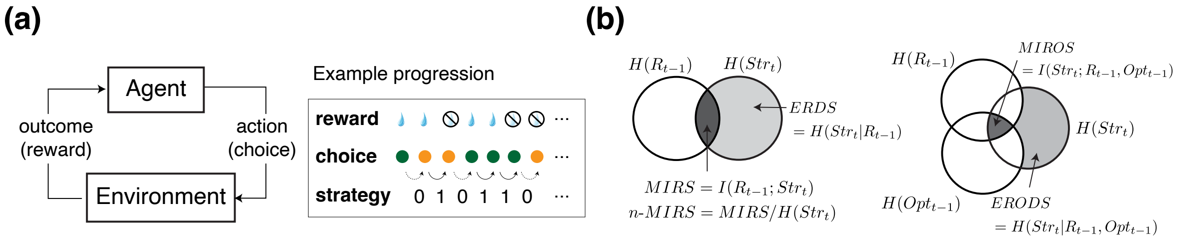

Figure 1. Probabilistic reversal learning task and information-theoretic framework for describing agent-environment interaction. (**a**) Illustration of agent-environment interaction and an example progression of task-related information used for computing behavioral information-theoretic metrics. (**b**) Left: illustration of conditional entropy and mutual information between previous reward (R_{t-1}) and current strategy (Str_t). Gray area represents conditional entropy of reward dependent strategy (ERDS $= H(Str_t \mid R_{t-1})$), and black area represents mutual information between strategy and previous reward (MIRS). n-MIRS denotes mutual information normalized by $H(Str)$. Right: similar metric defined on the joint combination of previous reward (R) and chosen option (Opt). Gray area represents conditional entropy of reward and option-dependent strategy (ERODS $= H(Str \mid R, Opt)$), and black area represents mutual information between reward outcome, choice options, and choice strategy (MIROS $= I(R, Opt; Str)$).

In general, for discrete random variables X and Y, the conditional entropy $H(Y|X)$ represents the remaining uncertainty in Y given the information about the variable X. Formally, it is defined as:

$$\begin{aligned} H(Y|X) &= \sum_{x \in X} P(x) \cdot H(Y|X = x) \\ &= -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)} \end{aligned} \tag{1}$$

Lower values of conditional entropy indicate that knowing the values of X reduces the uncertainty in Y, suggesting a strong dependence between the two variables.

Similarly, mutual information, denoted as $I(X;Y)$, quantifies the information shared between discrete variables X and Y. Higher values of $I(X;Y)$ indicate a greater dependency between variables, such that knowledge of Y would make X more predictable (less uncertain). This relationship is expressed using the difference between Shannon entropy for Y, $H(Y)$, and the conditional entropy, $H(Y|X)$:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \end{aligned} \tag{2}$$

Building on these general formulations, we define behavioral metrics to quantify uncertainty in the agent's choice strategy, specifically in terms of whether the agent "stays" with or "switches" from the previous choice option, given certain outcomes (**Figure 1a**). Specifically, we aim to quantify how uncertainty in choice strategy is reduced by certain task-related information: the previous reward outcome (R), and the previously chosen option (Opt). The resulting metrics include the conditional entropy of reward-dependent strategy (ERDS), the conditional entropy of option-dependent strategy (EODS), and the conditional entropy of reward and option-dependent strategy (ERODS). These are paired with mutual information metrics including mutual information between reward outcome and choice strategy (MIRS), mutual information between previous choice and choice strategy (MIOS), and mutual information between reward outcome, choice options, and choice strategy (MIROS).

More specifically, ERDS measures the influence of previous reward outcomes—reward vs. no reward (referred to as win vs. loss for simplicity)—on the uncertainty of the subsequent choice strategy (stay vs. switch), as follows:

$$\begin{aligned}\text{ERDS} &= H(Str_t \mid R_{t-1}) \\ &= H(Str) - I(R; Str) \\ &= -\sum_{R\in\{\text{win,loss}\}} \sum_{Str\in\{\text{stay,switch}\}} P(R, Str) \log_2 \frac{P(R, Str)}{P(R)}.\end{aligned} \tag{3}$$

where Str_t or Str denotes the choice strategy between two subsequent trials (stay = 1, switch = 0) , R_{t-1} is the reward outcome on the previous trial (reward or win = 1 and no reward or loss = 0), $P(R_{t-1}, Str_t)$ is the joint probability of choice strategy given a reward outcome on the previous trial, and $P(R)$ is the probability of reward. In the equation above, $H(Str)$ is the Shannon entropy of strategy, measuring the randomness in choice strategy in terms of stay or switch:

$$H(Str) = -\sum_{Str\in\{\text{stay,switch}\}} -P(Str) \log_2 P(Str) = (P(\text{stay}) \cdot \log_2 P(\text{stay}) + P(\text{switch}) \cdot \log_2 P(\text{switch})), \tag{4}$$

where $P(Str)$ is the probability of stay or switch ($P(stay) = 1 - P(switch)$). $I(R_{t-1}; Str_t)$, which we refer to as MIRS, is the mutual information between reward outcome and strategy, equal to:

$$\begin{aligned}\text{MIRS} &= I(R_{t-1}; Str_t) \\ &= \sum_{R\in\{\text{win,loss}\}} \sum_{Str\in\{\text{stay,switch}\}} P(R, Str) \log_2 \frac{P(R, Str)}{P(R)P(Str)}.\end{aligned} \tag{5}$$

Analogously, the conditional entropy of option-dependent strategy (EODS) represents the remaining uncertainty in the agent's strategy after accounting for the choice made on the previous trial. It is defined as the difference between $H(Str)$ and the mutual information between the previous choice and strategy (MIOS), as follows:

$$\begin{aligned}\text{EODS} &= H(Str_t \mid Opt_{t-1}) \\ &= H(Str) - I(Opt; Str) \\ &= -\sum_{Opt\in\{\text{better,worse}\}} \sum_{Str\in\{\text{stay,switch}\}} P(Opt, Str) \log_2 \frac{P(Opt, Str)}{P(Opt)},\end{aligned} \tag{6}$$

where Opt_{t-1} indicates the option chosen on the previous trial, with 1 indicating the better option and 0 indicating the worse option, as defined by the assigned reward probabilities, and P(Opt) is the probability of choosing the better option. $I(Opt_{t-1}; Str_t)$ denotes the mutual information between the chosen option and the agent's strategy, referred to as MIOS, and is calculated as follows:

$$\begin{aligned}\text{MIOS} &= I(Opt_{t-1}; Str_t) \\ &= \sum_{Opt\in\{\text{better,worse}\}} \sum_{Str\in\{\text{stay,switch}\}} P(Opt, Str) \log_2 \frac{P(Opt, Str)}{P(Opt)P(Str)}.\end{aligned} \tag{7}$$

To consider the combined effect of reward outcome and chosen option (i.e., winning or losing after choosing the better or worse option), we also considered a generalized metric that quantifies the combined effects of the two variables. Specifically, we define the conditional entropy of reward- and option-dependent strategy (ERODS), as follows:

$$\begin{aligned}\text{ERODS} &= H(Str_t \mid R_{t-1}, Opt_{t-1}) \\ &= H(Str) - I(R, Opt; Str) \\ &= -\sum_{R\in\{\text{win,loss}\}} \sum_{Opt\in\{\text{better,worse}\}} \sum_{Str\in\{\text{stay,switch}\}} P(R, Opt, Str) \log_2 \frac{P(R, Opt, Str)}{P(R, Opt)}.\end{aligned} \tag{8}$$

Here, $I(R, Opt; Str)$ denotes the mutual information between the combination of the previous reward and choice outcomes (R, Opt) and subsequent choice strategy (MIROS):

$$\begin{aligned}\text{MIROS} &= I(R_{t-1}, Opt_{t-1}; Str_t) \\ &= \sum_{R\in\{\text{win,loss}\}} \sum_{Opt\in\{\text{better,worse}\}} \sum_{Str\in\{\text{stay,switch}\}} P(R, Opt, Str) \log_2 \frac{P(Opt, Str)}{P(R, Opt)P(Str)}.\end{aligned} \tag{9}$$

2.2.1. Decomposition of Information-Theoretic Metrics

For the metrics described above, we also analyzed their decompositions into component values associated with each specific outcome of the conditioning variable. In the case of conditional entropy measures, this corresponds to the conditional entropy of Y given a specific value of $X = x$, weighted by the probability $P(x)$. For instance, ERDS can be decomposed into two components based on previous reward outcomes: $ERDS_+$ and $ERDS_-$, which reflect the uncertainty reduction following rewarded (win) and unrewarded (loss) trials, respectively [29]:

$$\begin{aligned}\text{ERDS}_+ &= H(Str_t \mid R_{t-1} = \text{win}) \cdot P(\text{win}) \\ &= -\left(P(\text{stay, win}) \cdot \log_2 \frac{P(\text{stay, win})}{P(\text{win})} + P(\text{switch, win}) \cdot \log_2 \frac{P(\text{switch, win})}{P(\text{win})}\right),\end{aligned} \tag{10}$$

$$\begin{aligned}\text{ERDS}_- &= H(Str_t \mid R_{t-1} = \text{loss}) \cdot P(\text{loss}) \\ &= -\left(P(\text{stay, loss}) \cdot \log_2 \frac{P(\text{stay, loss})}{P(\text{loss})} + P(\text{switch, loss}) \cdot \log_2 \frac{P(\text{switch, loss})}{P(\text{loss})}\right).\end{aligned} \tag{11}$$

This decomposition guarantees that $\text{ERDS}_+ + \text{ERDS}_- = \text{ERDS}$ (Equation (3)).

Similarly, EODS can be decomposed into $EODS_B$ and $EODS_W$ based on the choice of the better or worse option on the previous trial:

$$\begin{aligned}\text{EODS}_B &= H(Str_t \mid Opt_{t-1} = \text{better}) \cdot P(\text{better}) \\ &= -\left(P(\text{stay, better}) \cdot \log_2 \frac{P(\text{stay, better})}{P(\text{better})} + P(\text{switch, better}) \cdot \log_2 \frac{P(\text{switch, better})}{P(\text{better})}\right),\end{aligned} \tag{12}$$

$$\begin{aligned}\text{EODS}_W &= H(Str_t \mid Opt_{t-1} = \text{worse}) \cdot P(\text{worse}) \\ &= -\left(P(\text{stay}, \text{worse}) \cdot \log_2 \frac{P(\text{stay}, \text{worse})}{P(\text{worse})} + P(\text{switch}, \text{worse}) \cdot \log_2 \frac{P(\text{switch}, \text{worse})}{P(\text{worse})}\right).\end{aligned} \tag{13}$$

Finally, ERODS is decomposed into four components based on the combination of the previous reward and the choice option:

$$\begin{aligned}\text{ERODS}_{B+} &= H(Str_t \mid R_{t-1} = \text{win}, Opt_{t-1} = \textit{better}) \cdot P(\text{win,better}) \\ &= -\left(P(\text{stay}, \text{win}, \text{better}) \cdot \log_2 \frac{P(\text{stay}, \text{win}, \text{better})}{P(\text{win}, \text{better})}\right. \\ &\left.+ P(\text{switch}, \text{win}, \text{better}) \cdot \log_2 \frac{P(\text{switch}, \text{win}, \text{better})}{P(\text{win}, \text{better})}\right),\end{aligned} \tag{14}$$

$$\begin{aligned}\text{ERODS}_{B-} &= H(Str_t \mid R_{t-1} = \text{loss}, Opt_{t-1} = \textit{better}) \cdot P(\text{loss,better}) \\ &= -\left(P(\text{stay}, \text{loss}, \text{better}) \cdot \log_2 \frac{P(\text{stay}, \text{loss}, \text{better})}{P(\text{loss}, \text{better})}\right. \\ &\left.+ P(\text{switch}, \text{loss}, \text{better}) \cdot \log_2 \frac{P(\text{switch}, \text{loss}, \text{better})}{P(\text{loss}, \text{better})}\right),\end{aligned} \tag{15}$$

$$\begin{aligned}\text{ERODS}_{W+} &= H(Str_t \mid R_{t-1} = \text{win}, Opt_{t-1} = \textit{worse}) \cdot P(\text{win,worse}) \\ &= -\left(P(\text{stay}, \text{win}, \text{worse}) \cdot \log_2 \frac{P(\text{stay}, \text{win}, \text{worse})}{P(\text{win}, \text{worse})}\right. \\ &\left.+ P(\text{switch}, \text{win}, \text{worse}) \cdot \log_2 \frac{P(\text{switch}, \text{win}, \text{worse})}{P(\text{win}, \text{worse})}\right),\end{aligned} \tag{16}$$

$$\begin{aligned}\text{ERODS}_{W-} &= H(Str_t \mid R_{t-1} = \text{loss}, Opt_{t-1} = \text{worse}) \cdot P(\text{loss,worse}) \\ &= -\left(P(\text{stay}, \text{loss}, \text{worse}) \cdot \log_2 \frac{P(\text{stay}, \text{loss}, \text{worse})}{P(\text{loss}, \text{worse})}\right. \\ &\left.+ P(\text{switch}, \text{loss}, \text{worse}) \cdot \log_2 \frac{P(\text{switch}, \text{loss}, \text{worse})}{P(\text{loss}, \text{worse})}\right),\end{aligned} \tag{17}$$

where the subscripts refer to winning after choosing the better option ($B+$), losing after the better option ($B-$), winning after the worse option ($W+$), and losing after the worse option ($W-$). Note that ERODS can be alternatively decomposed based on either reward outcome or choice option alone, by adding the relevant components. For example, this metric can be decomposed based on whether the previous trial resulted in reward or no reward, as follows:

$$\begin{aligned}\text{ERODS}_{+} &= \text{ERODS}_{B+} + \text{ERODS}_{W+}, \\ \text{ERODS}_{-} &= \text{ERODS}_{B-} + \text{ERODS}_{W-}.\end{aligned} \tag{18}$$

This metric can also be decomposed based on whether the better or worse option was selected on the previous trial:

$$\begin{aligned}\text{ERODS}_{B} &= \text{ERODS}_{B+} + \text{ERODS}_{B-}, \\ \text{ERODS}_{W} &= \text{ERODS}_{W+} + \text{ERODS}_{W-}.\end{aligned} \tag{19}$$

For decomposing mutual information metrics, we utilize the general formulation given by:

$$\begin{aligned}I(X = x; Y) &= H(Y) - H(Y|X = x) \\ &= -\sum_{y \in Y} P(y) \cdot \log_2 P(y) + \sum_{y \in Y} P(y|x) \cdot \log_2 P(y|x),\end{aligned} \tag{20}$$

This quantity represents the information that a specific value $X = x$ provides about Y, and is known as pointwise mutual information. The expected value of pointwise mutual information over all values of X is equal to the mutual information between X and Y (Equation (2)):

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= \sum_{x\in X} P(x)H(Y) - \sum_{x\in X} P(x)H(Y \mid X = x) \\ &= \sum_{x\in X} P(x)\{H(Y) - H(Y \mid X = x)\} \\ &= \sum_{x\in X} P(x)I(X = x;Y) \end{aligned} \tag{21}$$

Unlike the mutual information $I(X;Y)$, which is always nonnegative, the pointwise mutual information for a specific event x, $I(X = x;Y)$, can take negative values when $H(Y) < H(Y|X = x)$ [65]. Conceptually, negative values indicate that the outcome x is misleading—or misinformative—about Y [66].

Using this definition, we compute two components of MIRS, MIRS_+ and MIRS_-, corresponding to previously rewarded and unrewarded outcome, respectively:

$$\begin{aligned} \text{MIRS}_+ &= I(R_{t-1} = \text{win}; Str_t) \cdot P(\text{win}) \\ &= \{H(Str) - H(Str_t \mid R_{t-1} = \text{win})\} \cdot P(\text{win}), \end{aligned} \tag{22}$$

$$\begin{aligned} \text{MIRS}_- &= I(R_{t-1} = \text{loss}; Str) \cdot P(\text{loss}) \\ &= \{H(Str) - H(Str_t \mid R_{t-1} = \text{loss})\} \cdot P(\text{loss}). \end{aligned} \tag{23}$$

These include the decomposition terms for conditional entropy (Equations (10)-11), which satisfy the identity ($\text{MIRS} = \text{MIRS}_+ + \text{MIRS}_-$). Similarly, the decompositions for mutual information between the option chosen on the previous trial and strategy, MIOS, are defined as:

$$\begin{aligned} \text{MIOS}_B &= I(Opt_{t-1} = \text{better}; Str_t) \cdot P(\text{better}) \\ &= \{H(Str) - H(Str_t \mid Opt_{t-1} = \text{better})\} \cdot P(\text{better}), \end{aligned} \tag{24}$$

$$\begin{aligned} \text{MIOS}_W &= I(Opt_{t-1} = \text{worse}; Str_t) \cdot P(\text{worse}) \\ &= \{H(Str) - H(Str_t \mid Opt_{t-1} = \text{worse})\} \cdot P(\text{worse}). \end{aligned} \tag{25}$$

These terms satisfy $\text{MIOS} = \text{MIOS}_B + \text{MIOS}_W$. Lastly, the decompositions for MIROS are given as:

$$\begin{aligned} \text{MIROS}_{B+} &= I(R_{t-1} = \text{win}, Opt_{t-1} = \text{better}; Str_t) \cdot P(\text{win,better}) \\ &= \{H(Str) - H(Str_t \mid R_{t-1} = \text{win}, Opt_{t-1} = \text{better})\} \cdot P(\text{win,better}), \end{aligned} \tag{26}$$

$$\begin{aligned} \text{MIROS}_{B-} &= I(R_{t-1} = \text{loss}, Opt_{t-1} = \text{better}; Str_t) \cdot P(\text{loss,better}) \\ &= \{H(Str) - H(Str_{t-1} \mid R_{t-1} = \text{loss}, Opt_{t-1} = \text{better})\} \cdot P(\text{loss,better}), \end{aligned} \tag{27}$$

$$\begin{aligned} \text{MIROS}_{W+} &= I(R_{t-1} = \text{win}, Opt_{t-1} = \text{worse}; Str_t) \cdot P(\text{win,worse}) \\ &= \{H(Str) - H(Str_t \mid R_{t-1} = \text{win}, Opt_{t-1} = \text{worse})\} \cdot P(\text{win,worse}), \end{aligned} \tag{28}$$

$$\begin{aligned} \text{MIROS}_{W-} &= I(R_{t-1} = \text{loss}, Opt_{t-1} = \text{worse}; Str_t) \cdot P(\text{loss,worse}) \\ &= \{H(Str) - H(Str_t \mid R_{t-1} = \text{loss}, Opt_{t-1} = \text{worse})\} \cdot P(\text{loss,worse}). \end{aligned} \tag{29}$$

These terms quantify the shared information between each instance of reward-option combination and the agent's strategy. Alternative decompositions can be computed similarly to ERODS as in Equations (18) and (19). Notably, these decompositions reveal how specific combinations of trial outcomes influence subsequent strategy, providing additional information beyond the MIRS and MIOS metrics.

2.2.2. Normalization of Mutual Information Metrics

The mutual information metrics defined thus far quantify the total amount of shared information between the random variables X and Y. However, one might be interested in the *fraction* of uncertainty in Y that is reduced by the knowledge of X, rather than the absolute amount of shared information [67–69]. In our setting, if the agent's strategy is largely deterministic (i.e., $H(Str)$ is low), then the mutual information will also be low due to small value of $H(Str)$ itself (e.g., see the area of gray circle specifying $H(Str)$ in **Figure 1b**). Therefore, to account for the baseline uncertainty in the strategy, we also considered normalized mutual information metrics, obtained by dividing the mutual information by $H(Str)$.

More specifically, we define normalized mutual information metrics for MIRS, MIOS, MIROS as follows:

$$\text{n-MIRS} = I(R_{t-1}; Str_t)/H(Str), \tag{30}$$

$$\text{n-MIOS} = I(Opt_{t-1}; Str_t)/H(Str), \tag{31}$$

$$\text{n-MIROS} = I(R_{t-1}, Opt_{t-1}; Str_t)/H(Str). \tag{32}$$

These metrics quantify the proportion of uncertainty in $H(Str)$ that is explained by each corresponding mutual information term. The normalized decompositions for each metric are similarly obtained by dividing each component by $H(Str)$. For example, normalized mutual information between reward and strategy is computed as:

$$\begin{aligned} \text{n-MIRS} &= \text{n-MIRS}_{+} + \text{n-MIRS}_{-}, \\ &= \frac{P(\text{win}) \cdot I(R_{t-1} = \text{win}; Str_t)}{H(Str)} + \frac{P(\text{loss}) \cdot I(R_{t-1} = \text{loss}; Str_t)}{H(Str)}, \end{aligned} \tag{33}$$

and analogously, normalized mutual information between choice option and strategy is defined as:

$$\begin{aligned} \text{n-MIOS} &= \text{n-MIOS}_{B} + \text{n-MIOS}_{W}, \\ &= \frac{P(\text{better}) \cdot I(Opt_{t-1} = \text{better}; Str_t)}{H(Str)} + \frac{P(\text{worse}) \cdot I(Opt_{t-1} = \text{worse}; Str_t)}{H(Str)}. \end{aligned} \tag{34}$$

Finally, the normalized mutual information between reward-choice combination and strategy can be computed as follows (dropping subscripts for trials for simplicity):

$$\begin{aligned} \text{n-MIROS} &= \text{n-MIROS}_{B+} + \text{n-MIROS}_{B-} + \text{n-MIOS}_{W+} + \text{n-MIOS}_{W+}, \\ &= \frac{P(\text{better,win}) \cdot I(R = \text{win}, Opt = \text{better}; Str)}{H(Str)} + \frac{P(\text{better,loss}) \cdot I(R = \text{loss}, Opt = \text{better}; Str)}{H(Str)} \\ &+ \frac{P(\text{worse,win}) \cdot I(R = \text{win}, Opt = \text{worse}; Str)}{H(Str)} + \frac{P(\text{worse,loss}) \cdot I(R = \text{loss}, Opt = \text{worse}; Str)}{H(Str)}. \end{aligned} \tag{35}$$

2.3. Reinforcement Learning Models and Simulations

We used multiple reinforcement learning (RL) models to simulate two variants of the probabilistic reversal learning task in order to illustrate that information-theoretic metrics can detect: (1) the differential learning rates for rewarded versus unrewarded trials, (2) changes in learning rates due to metaplasticity, (3) the influence of reward harvest rate, and (4) the use of alternative learning strategies in multi-dimensional reward environments.

For each task, we generated synthetic choice data using a range of model parameters and computed a set of relevant behavioral metrics from the resulting data. Below, we outline the general architecture of the RL models used, which are based on the standard Q-learning algorithm [22,25–27]. In this algorithm, the reward values of the two choice options, Q_A and Q_B, are updated on a trial-by-trial basis using reward feedback, and the difference between these values is used to make a choice (**Figure 2a**). Specifically, the value of the chosen option on a given trial t, Q_c, is updated according to the following rule:

$$Q_c(t+1) = Q_c(t) + \alpha(R(t) - Q_c(t)), \tag{36}$$

where $R(t)$ is the binary reward outcome (1 if rewarded, 0 if unrewarded), and α denotes the learning rate, which determines the magnitude of the value update by scaling the reward prediction error, $RPE = R(t) - Q_c(t)$. The value of the unchosen option remains unchanged. All value estimates are initialized at 0.5 for the first trial.

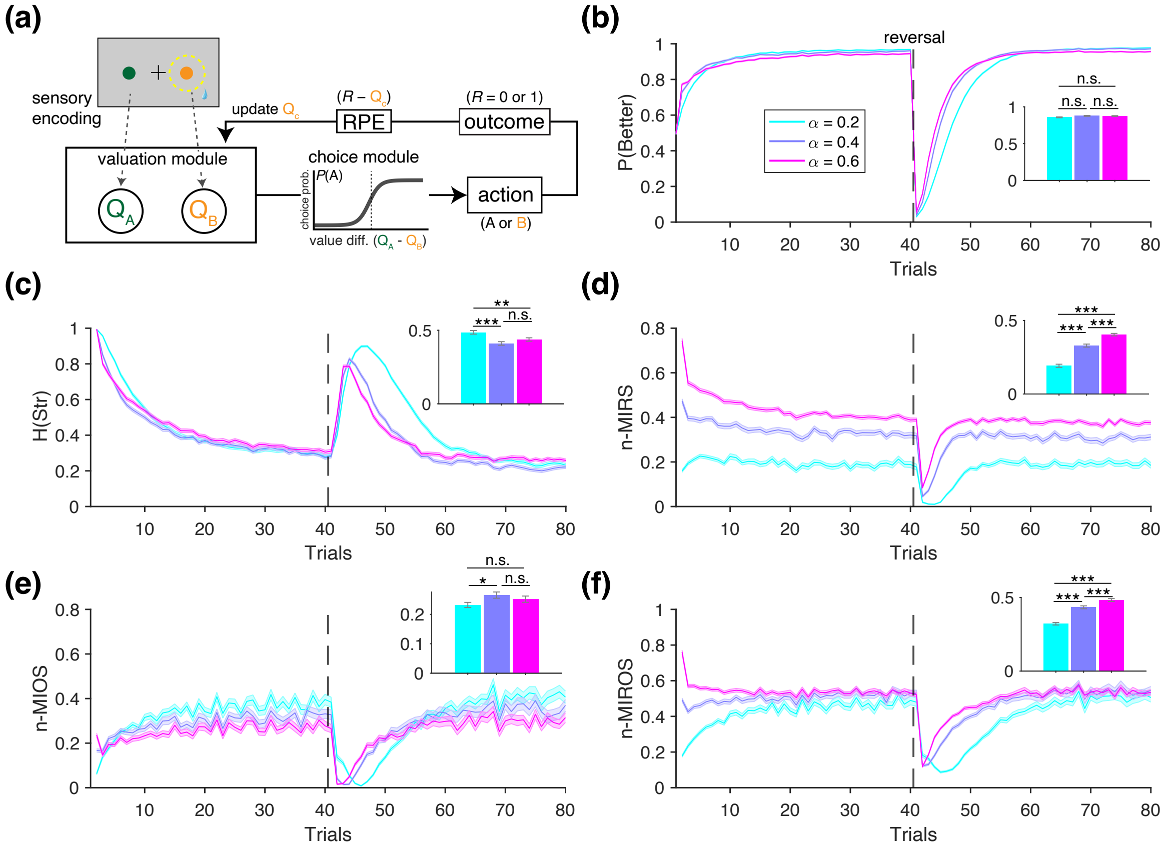

Figure 2. Schematic of an RL agent performing a probabilistic reversal learning task, with behavioral quantification using information-theoretic metrics. (**a**) Schematic of a standard RL algorithm that makes choices based on the value difference between the two options ($Q_A - Q_B$), using a softmax function to compute choice probabilities ($P(A)$). After each choice, the model updates the value estimate of the chosen option, Q_c, based on the difference between the reward outcome (R) and Q_c. This difference is known as the reward prediction error (RPE). (**b**) Plot shows average performance over time, defined as probability of choosing the better- rewarding option, shown separately for three different learning rates: $\alpha = 0.2, 0.4, 0.6$. Inset shows the average performance over 100 simulated blocks. 'n.s.' denotes no significant difference. (**c–f**) Plots show the averaged time course of the entropy of strategy ($H(Str)$; c), normalized mutual information between strategy and previous reward (n-MIRS; d), normalized mutual information between strategy and option (n-MIOS; e), and normalized mutual information between strategy and the previous reward/option combinations (n-MIROS; f). Asterisks indicate significant differences based on a two-sided rank sum test (*: $p < .05$; **: $p < .01$; ***: $p < .001$).

On each trial, the choice is determined stochastically based on the probability of selecting each option, P_j, computed using the softmax function as follows:

$$P_j(t) = \frac{e^{\beta Q_j(t)}}{e^{\beta Q_j(t)} + e^{\beta Q_k(t)}} = \frac{1}{1 + e^{-\beta(Q_j(t) - Q_k(t))}}, \tag{37}$$

where the indices j and k correspond to the two choice alternatives, and β is the inverse temperature parameter that controls the sensitivity to value differences (i.e., slope of the softmax function). We selected the softmax function because it is widely adopted in empirical research and follows naturally from normative rationality principles [70,71]. In its simplest form, the RL model has only two free parameters: α and β. Below, we describe the variants of the RL models used in each set of simulations.

2.3.1. Simulations of Positivity Bias

In this set of simulations, we examined the positivity bias—also known as the optimism bias—which refers to the tendency of an agent to learn more from feedback with positive valence than from feedback with negative valence [42–45]. This bias is often quantified by estimating separate learning rates for the outcomes that either are “better” or “worse” than expected, or equivalently, for positive and negative prediction errors, in the context of error-driven learning. In the case of binary reward outcomes, these correspond to the learning rates following rewarded and unrewarded trials, respectively. Therefore, the learning rule in Equation (36) can be generalized as:

$$Q_c(t+1) = \begin{cases} Q_c(t) + \alpha_+(R(t) - Q_c(t)), & \text{if } R(t) = 1 \\ Q_c(t) + \alpha_-(R(t) - Q_c(t)), & \text{if } R(t) = 0, \end{cases} \tag{38}$$

where α_+ and α_- denote the learning rates used to update the value of the chosen option on rewarded and unrewarded trials, respectively. Positivity bias is thus formally characterized by the condition where $\alpha_+ > \alpha_-$.

We used the above RL model to simulate the choice behavior in a probabilistic reversal learning task. Each block consisted of 80 trials with an 80/20% reward schedule, and a reversal in reward contingencies occurred after trial 40. To examine how positivity bias can be detected using information-theoretic metrics, we simulated choice behavior using different combinations of (α_+, α_-) values. We set the inverse temperature at $\beta = 10$, consistent with values observed in previous experimental studies [42,72]. To explore a plausible range of learning rates while avoiding extreme values, we varied each learning rate $\alpha_i \in [0.1, 0.9]$ in increments of 0.05, resulting in a 17-by-17 grid of (α_+, α_-) combinations. For each point in this parameter space, choice behavior was simulated over 10,000 blocks.

To test whether the information-theoretic measures contain sufficient information to determine the presence of positivity bias ($\alpha_+ > \alpha_-$), we utilized a linear decoder based on logistic regression, implemented using the *fitclinear* function in MATLAB). Decoders were trained on metrics computed from the choice behavior of training samples and tested on a separate held-out set, with the goal of predicting whether the underlying RL model used to generate the data had $\alpha_+ > \alpha_-$.

More specifically, in each decoding experiment, we randomly sampled RL agents, each assigned a true (α_+, α_-) pair drawn independently from uniform distributions over $\alpha_i \in [0.1, 0.9]$ (**Figure 3e**). Twenty RL agents were assigned to each group, labeled as either "optimistic" ($\alpha_+ > \alpha_-$) or "pessimistic" ($\alpha_+ < \alpha_-$). For each agent, every simulated block (out of total of 100) used a distinct (α_+, α_-) pair sampled from gaussian distributions ($\sigma = 0.03$) centered on that agent’s true (α_+, α_-) values. For each block, we obtained theoretical averages of the information-theoretic metrics by averaging across 10 repeated simulations. To evaluate decoding performance, we employed a leave-one-out procedure at the agent level: the decoder was trained on the information-theoretic metrics computed from the choice behavior of all agents except the one being tested. The training data were balanced to include an equal number of optimistic and pessimistic agents. Decoding accuracy was averaged across 100 independent decoding experiments, each using a unique set of RL agents.

Finally, to benchmark our decoder against behavioral features beyond the information-theoretic metrics, we trained a separate decoder on the coefficients of a logistic regression predicting choice from past choice and reward history. More specifically, we fit the following logistic regression model [30] to the same segment of choice data:

$$\log\left(\frac{P(C(t)=j)}{P(C(t)=k)}\right) = \beta_0 + \sum_{i=1}^{5} \beta_i^R (R_j(t-i) - R_k(t-i)) + \sum_{i=1}^{5} \beta_i^U (U_j(t-i) - U_k(t-j)), \tag{39}$$

where $C(t)$ indicates the choice made on trial t, j and k index the two choice options, R_i indicates a rewarded choice (1 if rewarded, 0 otherwise), and U_i denotes an unrewarded choice (1 if unrewarded, 0 otherwise). We used the resulting ten regression coefficients (β^R and β^U) to decode whether a given agent is optimistic or pessimistic.

2.3.2. Simulations of Reward-Dependent Metaplasticity

In this set of simulations, we examined whether adjustments in the learning rates—predicted by reward-dependent metaplasticity [46,47]—can be detected using information-theoretic metrics. Importantly, reward-dependent metaplasticity enables dynamic adjustments of synaptic plasticity over time and can account for stimulus- or action-specific learning rates, including asymmetries such as positivity bias. Using the same probabilistic reversal learning task with an 80/20% reward schedule as in the positivity bias simulations, we compared the choice behavior of the metaplastic model to that of the 'plastic' model, which is equivalent to the standard RL model described earlier. Below, we briefly describe the metaplastic model (see [46] for more details).

Importantly, the standard RL algorithm can be implemented through binary synapses that encode and update the value of each choice option through reward-dependent plasticity: transitioning from a "weak" to a "strong" state following reward (reward-dependent potentiation), and from a "strong" to a "weak" state following no reward (reward-dependent depression) [47,73,74] (**Figure 4a**). Importantly, the proportion of synapses in the "strong" state provides an estimate of the value of a given choice option (Q_j), as this quantity increases and decreases following reward and no reward, respectively [73]. We assume that only the synapses associated with the chosen option undergo state transitions according to the reward outcome (reward-dependent plasticity), while those associated with the unchosen option remain unchanged—consistent with the assumption of standard Q-learning model.

The metaplastic model generalizes the above reward-dependent plasticity mechanism by introducing multiple meta-states for each level of synaptic efficacy, with synapses occupying deeper meta-states being more resistant—or stable—with respect to future changes (**Figure 4b**). More specifically, synapses in the metaplastic model undergo both plastic and metaplastic transitions. During plastic transitions, synaptic efficacy shifts between "weak" and "strong" states. During metaplastic transitions, synaptic efficacy remains unchanged, but the stability is modified as the synapse transitions to a deeper or shallower meta-state (**Figure 4b**).

Formally, higher stability of deeper meta-states is captured by decreasing transition probabilities q_i for the i^{th} meta-state, governed by a power law as follows:

$$q_i = q_1^{\frac{i(m-2)+1}{m-1}} \quad \text{for } 2 \le i \le m, \tag{40}$$

where m is the number of meta-states, q_1 is the (baseline) transition probability between the most unstable weak and most unstable strong meta-states, and q_i is the transition probability from weak (or strong) meta-state $i+1$ to the most unstable strong (or weak) state in response to positive (or negative) reward feedback (reward vs. no reward), as indicated by the diagonal arrows in **Figure 4b**.

Additionally, metaplastic synapses can undergo transitions that do not alter their synaptic efficacy (vertical arrows in **Figure 4b**), with the probability of such transitions decreasing for deeper meta-states:

$$p_i = {p_1}^i \quad \text{for } 2 \le i \le m-1, \tag{41}$$

where p_i indicate the transition probability between the i^{th} and the $(i+1)^{th}$ meta-states. After positive reward feedback (potentiation events), "weak" synapses undergo transition toward less stable meta-states while "strong" synapses transition toward more stable, deeper meta-states. Conversely, after negative reward feedback (depression events) "weak" synapses undergo transition toward more stable, deeper meta-states while "strong" synapses transition toward less stable meta-states (vertical arrows in **Figure 4b**).

The value of each option in the metaplastic model is computed by summing over the fractions of synapses in the strong states:

$$Q_j(t) = \sum_{i=1}^{m} S_{j,i}(t), \tag{42}$$

where j indexes the choice option, and $S_{j,i}(t)$ indicates the fraction of synapses in the i^{th} strong meta-state on a given trial.

Because the metaplastic model includes multiple meta-states with different transition rates, its update in response to reward feedback—and thus its rate of learning—changes over time. This dynamic can be quantified using two "effective" learning rates computed on each trial. The effective learning rate following reward (α_+^{eff}) is defined as the fraction of synapses that transition from weak to strong states, while the effective learning rate following no reward (α_-^{eff}) reflects the fraction transitioning from strong to weak states, as follows:

$$\begin{aligned} \alpha_+^{eff}(t) &= \frac{Q_c(t+1) - Q_c(t)}{R(t) - Q_c(t)} \quad \text{if } R(t) = 1, \\ \alpha_-^{eff}(t) &= \frac{Q_c(t+1) - Q_c(t)}{R(t) - Q_c(t)} \quad \text{if } R(t) = 0, \end{aligned} \tag{43}$$

where the numerator indicates the change in the overall value of the chosen option, while the denominator corresponds to the prediction error.

To simulate choice behavior, we set $\alpha = 0.3$ and $\beta = 10$ for the plastic model with a single learning rate. For the metaplastic model, we used the model with $m = 4$ meta-states and set $p_1 = 0.4$ for the baseline meta-transition probability. We set the transition probability $q_1 = 0.516$, which yields an initial effective learning rate of 0.3 on the first trial of the block (**Figure 4c**). We also tested the choice behavior of a plastic model with differential learning rates for rewarded and unrewarded outcomes ($\alpha_+ \neq \alpha_-$). These learning rates were estimated by fitting the plastic model to the choice behavior of the metaplastic model on each block via maximum-likelihood estimation (optimized with *fmincon* function in MATLAB).

Similar to the positivity bias simulations, we used a linear decoder to test whether the emergence of positivity bias ($\alpha_+ > \alpha_-$)—driven by metaplasticity over time—can be detected using the information-theoretic metrics. The training procedure involved randomly drawing a set of (α_+, α_-) values from the grid of [0.2, 0.4] in a step size of 0.01, under the assumption that the approximate range of α value is known ($\alpha = 0.3$ in our case). Each parameter space was labeled as either "optimistic" ($\alpha_+ > \alpha_-$) or "pessimistic" ($\alpha_+ < \alpha_-$). The points with $\alpha_+ = \alpha_-$ were dropped from the training set. We then computed the decoder's posterior probability on held-out samples, which were generated from the choice behavior of the three models: (1) a plastic model with a single learning rate ($\alpha_+ = \alpha_- = \alpha$), (2) the metaplastic model, and (3) a plastic model with separate learning rates ($\alpha_+ \neq \alpha_-$) estimated from the choice behavior of the metaplastic model.

2.3.3. Simulations of Reward Harvest Rate Effects on Behavior

In this set of simulations, we examined whether the influence of the overall reward rate on a long time scale—referred to as reward harvest rate—can be detected using the information-theoretic metrics. We used the same probabilistic reversal learning task as in the positivity bias simulations and tested two RL models for generating choice behavior: (1) the standard RL model with a single learning rate, as described earlier; and (2) an augmented RL model that included an additional adjustment mechanism based on reward harvest rate (RL with RHR model; **Figure 5a**). The latter model served as the ground truth for a potential pathway through which reward harvest rate could influence choice strategy—specifically, by increasing the tendency to win-stay and lose-switch as the harvest reward rate increases. More specifically, the augmented model tracks a third variable H, representing the overall reward harvest rate independent of any specific choice option, and updates it as follows:

$$H(t+1) = H(t) + \alpha_H (R(t) - H(t)), \tag{44}$$

where α_H is the update rate for H. Therefore, H provides an exponentially weighted moving average of harvested rewards. For simplicity, we set $\alpha_H = \alpha = 0.3$ in the simulations. In practice, although α_H can be directly optimized to the data, one can equally justify choosing an arbitrary α_H for behavioral analysis to avoid modeling assumptions and obtain model-independent measures. To incorporate a win-stay/lose-switch bias that is modulated by the reward harvest rate H, the choice probability in this model is computed as follows:

$$P_j(t) = \frac{1}{1 + e^{-\left[\beta(Q_j(t) - Q_k(t)) + \beta_H Bias_j(t)\right]}}, \tag{45}$$

where β_H is a free parameter controlling the overall influence of the win-stay/lose-switch bias term for option j, denoted as $Bias$. This bias term is calculated as follows:

$$Bias_j(t) = C_j(t-1) \cdot \big(R(t-1) - H(t-1)\big), \tag{46}$$

where C_j indicates whether option j was chosen in the previous trial (1 if chosen, -1 otherwise), and the term $R - H$ dynamically adjusts the agent's tendency to repeat (after reward) or switch (after no reward) based on how much the recent reward outcome differs from reward harvest rate H. In our simulations, we set $\beta = 10, \beta_H = 2$.

To quantify the effect of reward harvest rate on choice strategy, we introduced an additional mutual information metric, referred to as MIRHS, which captures the joint influence of previous reward outcome and the "reward harvest state" (as defined by reward rate) on the agent's strategy, as follows:

$$\begin{aligned} \text{MIRHS} &= I(R_{t-1}, RHS_{t-1}; Str_t) \\ &= \sum_{R \in \{\text{win,loss}\}} \sum_{RHS \in \{\text{high,low}\}} \sum_{Str \in \{\text{stay,switch}\}} P(R, RHS, Str) \log_2 \frac{P(RHS, Str)}{P(R, RHS) P(Str)}, \end{aligned} \tag{47}$$

where RHS denotes the reward harvest state—a binary, discretized variable indicating whether the reward harvest rate H on a given (previous) trial was classified as "high" or "low", based on the median split within the current block (**Figure 5b**) [35]. We also used a normalized version of this metric, n-MIRHS $= I(R_{t-1}, RHS_{t-1}; Str_t)/H(Str_t)$, which is decomposed as follows (dropping subscripts for trials for simplicity):

$$\begin{aligned} \text{n-MIRHS} = {} & \text{n-MIRHS}_{high+} + \text{n-MIRHS}_{high-} + \text{n-MIRHS}_{low+} + \text{n-MIRHS}_{low+}, \\ & = \frac{P(\text{high,win}) \cdot I(R = \text{win}, RHS = \text{high}; Str)}{H(Str)} + \frac{P(\text{high,loss}) \cdot I(R = \text{loss}, RHS = \text{high}; Str)}{H(Str)} \\ & + \frac{P(\text{low,win}) \cdot I(R = \text{win}, RHS = \text{low}; Str)}{H(Str)} + \frac{P(\text{low,loss}) \cdot I(R = \text{loss}, RHS = \text{low}; Str)}{H(Str)}, \end{aligned} \tag{48}$$

where the subscripts of the metric refer to winning during a high RHS state ($high+$), losing during a high RHS state ($high-$), winning during a low RHS state ($low+$), and losing during a low RHS state ($low-$). These metrics therefore separate the influence of immediate, reward outcome from that of the reward harvest state. To isolate the overall effect of RHS, we also computed an alternative decompositions as follows:

$$\text{MIRHS}_{high} = \text{MIRHS}_{high+} + \text{MIRHS}_{high-}, \tag{49}$$

$$\text{MIRHS}_{low} = \text{MIRHS}_{low+} + \text{MIRHS}_{low-}, \tag{50}$$

2.3.4. Simulations of Learning in Multidimensional Reward Environments

To investigate the presence of alternative learning strategies in the agent's behavior, we conducted a final set of simulations using a variant of the probabilistic reversal learning task in which different attributes of the choice options predicted reward outcomes at different time points.

In this task, at any given point within a block of trials, only one of the two choice attributes—the shape and the color of the stimuli—is predictive of reward probabilities (80% vs 20%). Initially, the reward schedule for a given block is assigned to the color of the stimuli regardless of their shape, with higher reward probability assigned to either green or orange objects. Therefore, unlike the reversal learning task used in the previous simulations, this task involves two types of reversals: (i) reversal in the reward probability, as previously considered, and (ii) reversal in the feature predictive of reward (**Figure 6a**). Therefore, in a *Color-to-Color* (or similarly *Shape-to-Shape*) block, the reward contingency was reversed between the better and worse values of the same attribute—for example, between two colors (**Figure 6c**). In contrast, in a *Color-to-Shape* block, the predictive attribute switched, and reward became associated with the shape of the object (e.g., triangle = 80% and square = 20%), regardless of its color (**Figure 6d**). The identity of the block type was unknown to the agent, and the two block types were randomly interleaved throughout the session. The Agent therefore had to adapt their learning and choice strategies solely based on the reward feedback they received.

To capture the strategies that more realistic agents might adopt, we used a generalized RL model that dynamically arbitrates between competing learning strategies. To that end, we used a variant of the model introduced by Woo et al. [38], which simultaneously tracks the values of multiple choice attributes (**Figure 6b**). In this model, the value estimates for each attribute are updated simultaneously as follows:

$$Q_{c,i}(t+1) = Q_{c,i}(t) + \alpha(R(t) - Q_{c,i}(t)), \tag{51}$$

$$Q_{u,i}(t+1) = (1-\gamma)Q_{u,i}(t), \tag{52}$$

where i indexes the choice attribute (Color or Shape), and c indexes the chosen option within each attribute (i.e., Color $\in \{green, orange\}$ and Shape $\in \{triangle, square\}$). The index u refers to the unchosen option within each attribute, and γ denotes the decay rate of value for unchosen options. Note that this decay mechanism is biologically motivated and included to account for the increased number of choice features in this model, reflecting the brain's limited capacity to retain the memory of all value estimates. To generate a choice, the model first computes the overall value, V, of the two choice alternatives through a linear weighted combination of the attributes, as follows:

$$V_j(t) = Q_{Color,j} \times \omega(t) + Q_{Shape,j} \times (1 - \omega(t)), \tag{53}$$

where $\omega(t)$ is the arbitration weight on trial t specifying the relative contribution of color attribute to overall value, and the index j denotes the choice options to the left and right of the screen (for example, if the leftward option is a green triangle, $V_{left}(t) = Q_{green}(t)\omega(t) + Q_{triangle}(t)(1 - \omega(t))$). The difference in overall values V is then used to compute choice probability, similar to Equation (37). It is important to note that linear combination of feature values to an overall value is used primarily for simplicity in describing the model and does not necessarily imply that such an integrated value is explicitly constructed. In practice, it is more plausible that the value of each attribute is compared directly—albeit with different weights—when making choices [75].

To dynamically control the arbitration weight between competing strategies, we considered the following update rule:

$$\omega(t+1) = \begin{cases} \omega(t) + \alpha_\omega|\Delta Rel(t)|(1-\omega(t)), & \text{if } \Delta Rel(t) > 0 \\ \omega(t) + \alpha_\omega|\Delta Rel(t)|(0-\omega(t)), & \text{if } \Delta Rel(t) < 0, \end{cases} \tag{54}$$

where ΔRel specifies the difference in the reliability between color and shape attributes in predicting rewards. Intuitively, the model assigns higher weight to the attribute which is estimated to have higher reliability for the given trial. We defined reliability based on the absolute value of the reward prediction error (RPE) for each attribute. Specifically, the reliability of an attribute is defined as inversely proportional to the magnitude of its RPE—a lower RPE (i.e., less surprise) indicates higher reliability. Based on this definition, the reliability difference ΔRel is given by:

$$\begin{aligned} \Delta Rel(t) &= |RPE_{Shape}| - |RPE_{Color}|, \\ &= |R(t) - Q_{c,Shape}| - |R(t) - Q_{c,Color}|, \end{aligned} \tag{55}$$

where RPE_{Color} and RPE_{Shape} refer to the reward prediction error between actual and predicted reward outcomes based on color or shape attribute, respectively. For example, if the color attribute yields a smaller RPE magnitude than the shape attribute, then $\Delta Rel > 0$, and the model increases ω to bias decision making toward the color attribute on the next trial.

We considered three distinct decision-making strategies, each implemented as a special case of the arbitration model described above. In the first case, the model fixed $\omega = 1$ for all trials, representing an agent who relies exclusively on the color attribute. In the second case, $\omega = 0.5$ was fixed, modeling an agent who assigns equal weight to both color and shape attributes. In the third case, the full arbitration mechanism was implemented, with the agent dynamically adjusting ω based on the relative reliability of the two attributes in predicting reward. For this model, ω was initialized at 0.5 on the first trial. For the full model, we used the following parameter values: $\alpha = 0.4, \beta = 10, \gamma = 0.2, \alpha_\omega = 0.4$.

To quantify the relative dominance of color-based versus shape-based learning strategies, we computed the conditional entropy of reward-dependent strategy (ERDS) separately for each attribute. Specifically, we defined stay/switch strategies separately for color and shape—denoted Str_{Color} and Str_{Shape}, respectively—based on whether the agent repeats or switches their choice option with respect to each attribute after reward feedback. The corresponding ERDS measure for each attribute was then defined as follows:

$$\begin{aligned} \text{ERDS}_{Color} &= H(Str_{Color}|R), \\ \text{ERDS}_{Shape} &= H(Str_{Shape}|R), \end{aligned} \tag{56}$$

To quantify the relative dominance of strategies in response to reward feedback, we computed the difference $\Delta ERDS = ERDS_{Shape} - ERDS_{Color}$. A higher value of $\Delta ERDS$ indicates greater reliance on a color-based choice strategy.

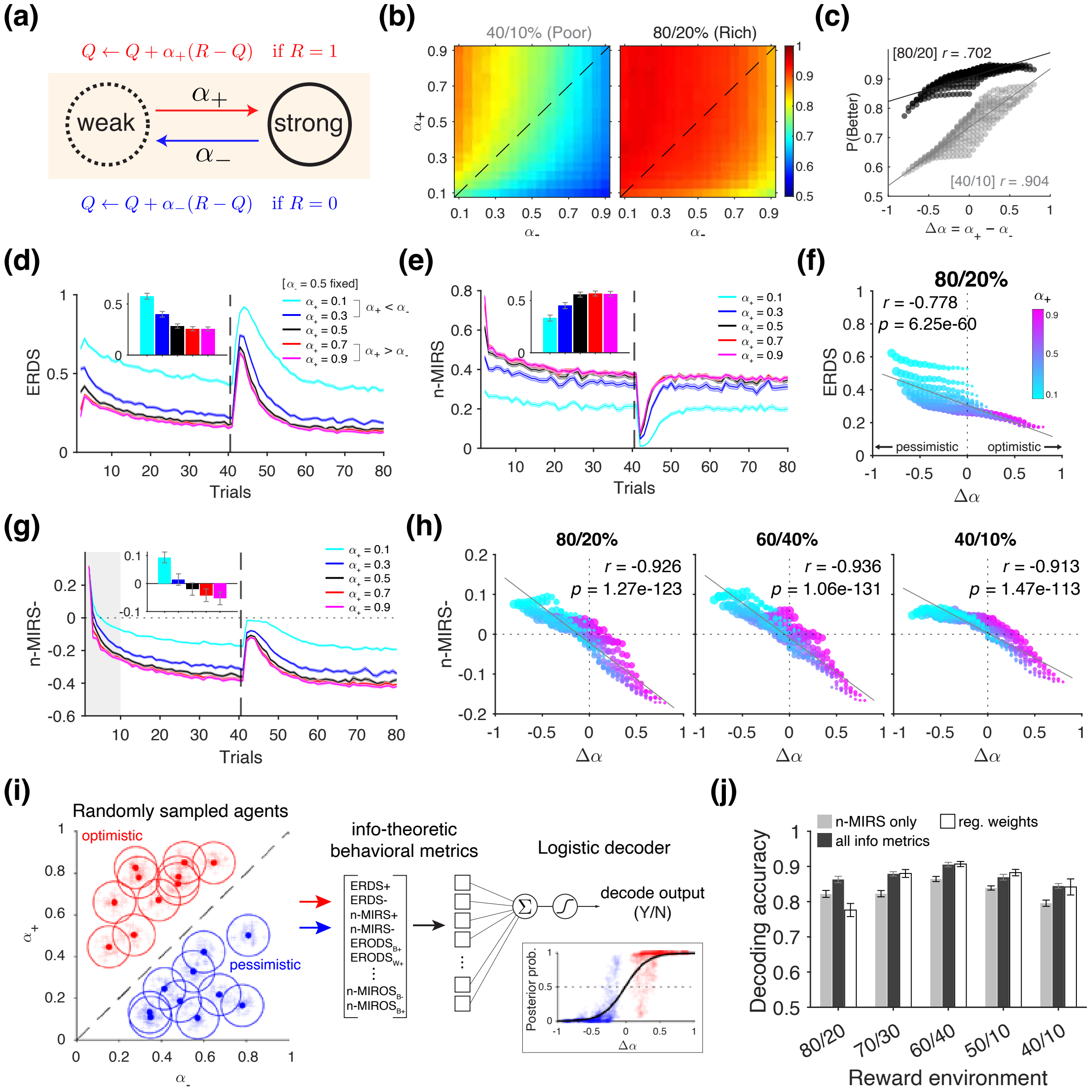

Figure 3. Behavioral signatures of positivity bias and its decoding using information-theoretic metrics. (**a**) Illustration of value updates following rewarded ($R = 1$) and unrewarded ($R = 0$) outcomes, with corresponding differential learning rates α_+ and α_-, respectively. (**b**) Behavioral signature of positivity bias. Heatmaps show the average performance for each combination of (α_+, α_-) in two example reward schedules. (**c**) Correlations between performance and the difference in the learning rates ($\Delta\alpha = \alpha_+ - \alpha_-$), shown separately for two reward schedules. (**d, e**) Example time courses of ERDS (d) and n-MIRS (e) for five example RL agents with different learning rates. Red/magenta lines indicate "optimistic" agents, blue/cyan lines indicates "pessimistic" agents, and black corresponds to the neutral agent. Insets show metric computed over the whole block. (**f**) Correlation between $\Delta\alpha$ and ERDS metric in the 80/20% environment. (**g**) Example time courses for the decomposition of n-MIRS after negative reward feedback (n-MIRS-). The sign of n-MIRS-, computed from the first 10 trials (inset), is predictive of positivity bias. (**h**) Correlations between $\Delta\alpha$ and n-MIRS- metric, shown separately for three example reward environments. (**i**) Illustration of sampling and decoding procedure to predict the sign of $\Delta\alpha$ using information-theoretic metrics. For each RL agent in the two groups—"optimistic" with $\Delta\alpha > 0$ or "pessimistic" with $\Delta\alpha < 0$)—a mean value of (α_+, α_-) was randomly drawn. Learning rates for each session were then randomly sampled from Gaussian distributions centered on these means. Behavioral metrics were computed from the resulting choice behavior in each session and used to train a binary decoder. The inset on the right shows the predicted posterior probability generated by cross-validated decoders for each true value of $\Delta\alpha$, with a sigmoid curve fitted to the data. (**j**) Cross-validated decoding accuracy from decoders trained on the information-theoretic metrics—n-MIRS- metric only (gray) or all available metrics (black)—and logistic regression weights (white empty bars), separately for different reward environments.

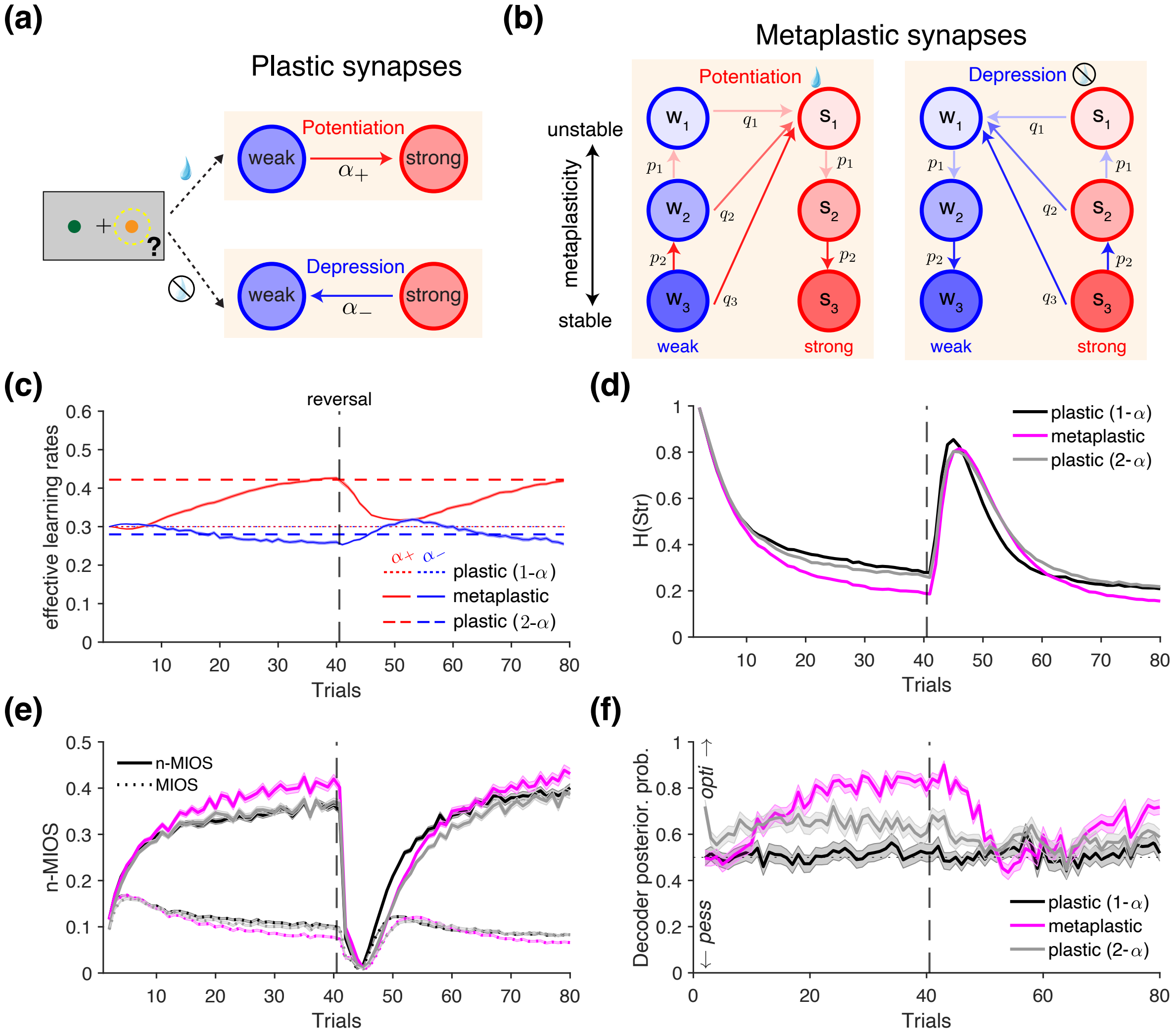

Figure 4. Comparison of plastic and metaplastic models and their distinct behavior revealed by information-theoretic metrics. (**a**) Schematic of a biological implementation of a standard RL model with binary synapses that undergo stochastic reward-dependent plasticity. In this model, "weak" synapses transition to the "strong" synaptic state (i.e., are potentiated) with probability α_+ following reward, whereas "strong" synapses transition to the "weak" state (i.e., are depressed) in the absence of reward. (**b**) Schematic of synapses with reward-dependent metaplasticity, consisting of three meta-states with increasing levels of stability ($i = 1, 2, 3$), each associated with one of the two levels of synaptic strengths ("weak" w or "strong" s). Transition probabilities q_i govern changes from the i^{th} meta-state to the most unstable meta-state of opposite efficacy, with $q_1 > q_2 > q_3$. Transition probabilities p_i govern transitions between meta-states of the same efficacy, and are also higher for less stable states (i.e., $p_1 > p_2$). (**c**) Averaged trajectories of "effective learning rates" for the metaplastic model (solid lines), a plastic model with a single learning rate (dotted lines; $\alpha = 0.3$), a the plastic model with differential learning rates estimated from the choice behavior of the metaplastic model (dashed lines). Red and blue correspond to the effective learning rates on rewarded and unrewarded trials, respectively. (**d**) Averaged trajectory of $H(Str)$, shown separately for the three models as indicated in the legend. (**e**) Averaged trajectory of normalized mutual information between previously chosen option and strategy (n-MIOS, solid lines), separately for each of the three model. MIOS before normalization by $H(Str)$ are plotted for reference (dotted lines). (**f**) Posterior probability that a given set of metrics was classified as originating from an "optimistic" (*opti*) vs. a "pessimistic" (*pess*) agent, based on binary decoders applied to the choice behavior of the three models.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2025 doi:10.20944/preprints202509.0001.v1

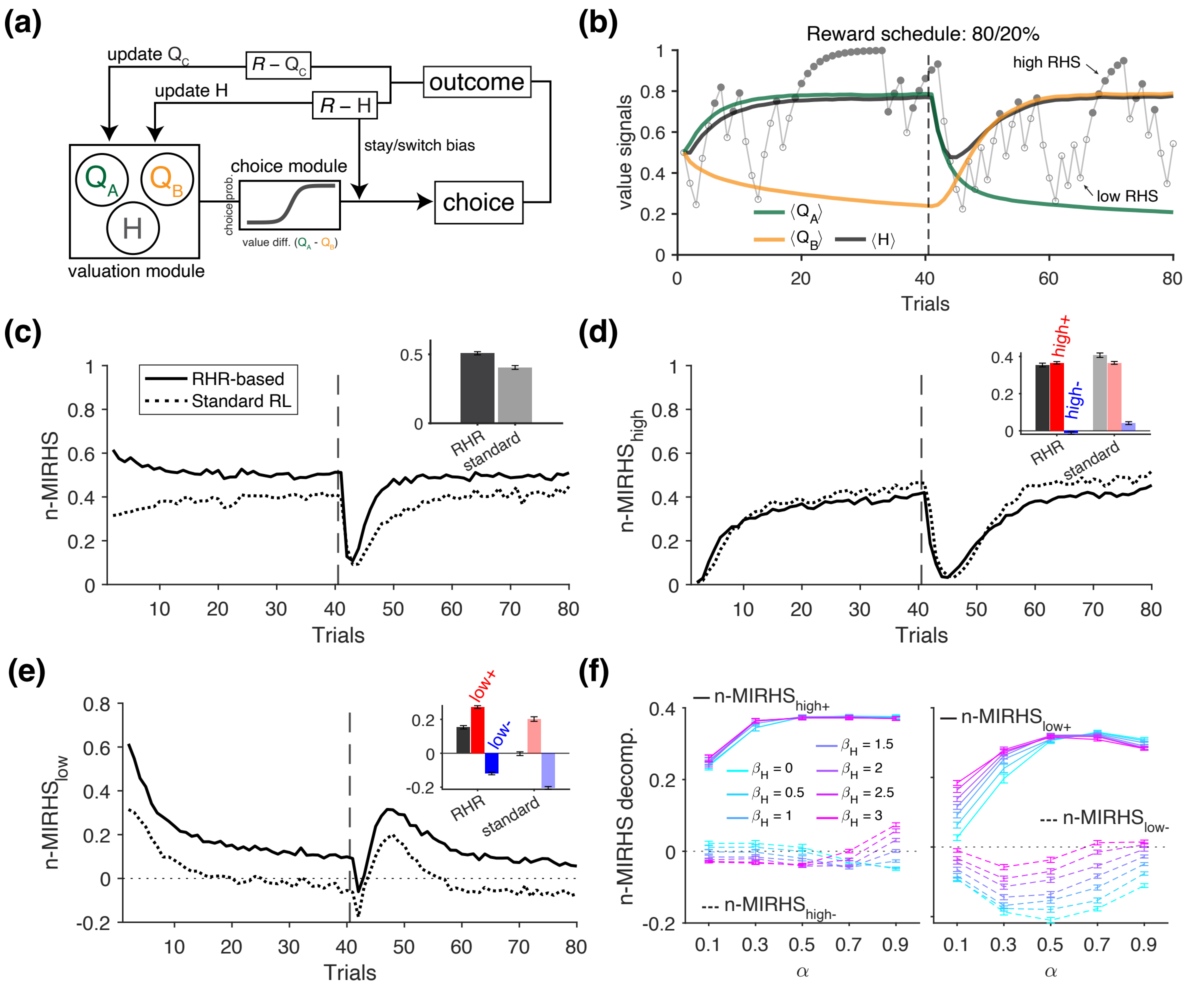

Figure 5. Detecting the modulation of choice strategy by reward harvest rate. (**a**) Schematic of an RL model that tracks reward harvest rate to modulate choice strategy. In this model, the difference between the immediate reward (R) and reward harvest rate (H) is used to directly influence win-stay and lose-switch behavior. (**b**) Averaged time course of the model estimated values (Q_i's in green and orange) and reward harvest rate (H in black) during the PRL task with a reversal on trial 40. The gray trace in the background depicts the time course of H from an example session. Filled (empty) circles represent trials with high (low) reward harvest states (*RHS*), determined via a median split of H values across the session. Simulation parameters: $\alpha = 0.3$, $\beta = 10$, $\beta_H = 2$. (**c**) Averaged time course of normalized mutual information between strategy and reward/RHS combinations (n-MIRHS), shown for models with and without the reward harvest rate modulation. (**d**) Decomposition of the n-MIRHS metric for trials in the high RHS state. Inset shows further decompositions into previously rewarded (*high+*, red) and unrewarded (*high-*, blue) trials. (**e**) Decomposition of the n-MIRHS metric for trials in the low RHS state. Inset shows further decompositions into previously rewarded (*low+*, red) and unrewarded (*low-*, blue) trials. (**f**) Plotted are the mean simulated n-MIRHS decompositions based on high (left panel) or low (right panel) reward state, following rewarded (solid) or unrewarded (dashed lines) trials, as a function of α (X-axis) and β_H (indicated by colors). The influence of β_H was most pronounced in the information gained from no reward in the low RHS.

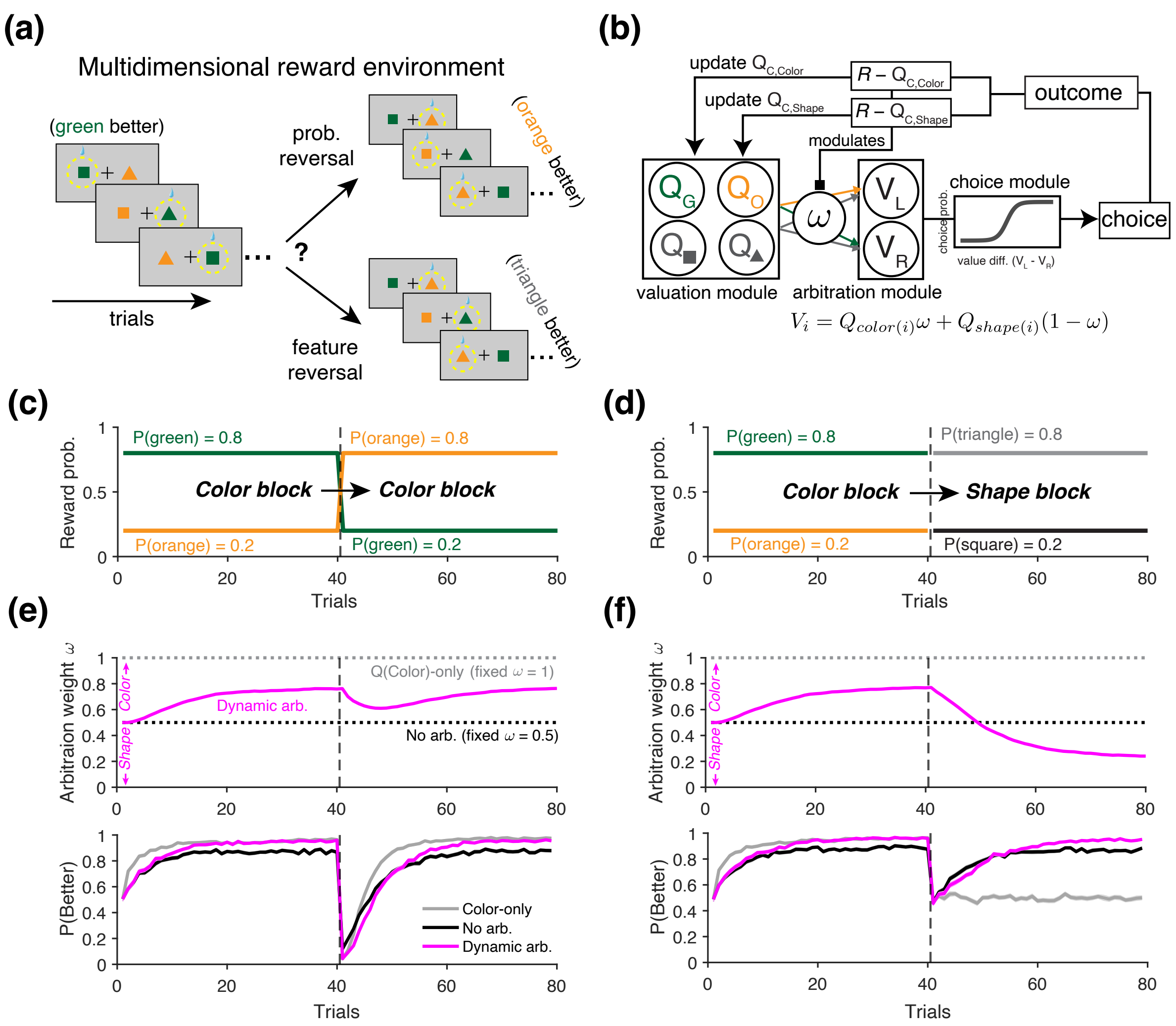

Figure 6. Schematic of a multidimensional reward learning task and RL models used to perform it. (**a**) Illustration of example trials from the multidimensional reversal learning task. Reversals can occur either in the reward probabilities associated with two options (e.g., switching from green to orange as the better option), as in the standard probabilistic reversal learning task, or in the feature dimension predictive of rewards (e.g., from color to shape). (**b**) Schematic of the dynamic arbitration model that simultaneously tracks the values of individual color and shape features (Q). These feature values are then combined to compute the overall value (V) of the left and right options. The arbitration weight (ω) is dynamically adjusted based on the reliability of the chosen color and shape values, estimated from their respective reward prediction errors. (**c**) Illustration of a reversal where the better option reverses from the green to the orange object (Color-to-Color reversal), regardless of shape. (**d**) Illustration of a reversal where the better option reverses from the green to triangle object (Color-to-Shape, feature reversal). (**e**) Averaged time courses of arbitration weight (ω, top) and task performance (bottom) during Color-to-Color reversal blocks for the three models. The dynamic arbitration model (*Dynamic arb.* gradually increases its weighting of color. The model with no dynamic arbitration (*No arb.* maintains a fixed weight of $\omega = 0.5$, equally combining color and shape values. The *Color-only* model does not track shape values and therefore, ω remains fixed at 1. (**f**) Same plots as in (e) but for Color-to-Shape (feature reversal) blocks.

3. Results

In the following sections, we illustrate the utility of information-theoretic metrics for identifying distinct learning and decision-making mechanisms using simulated data described above. We begin by presenting example behavior of an RL agent with a single learning rate, shown for three values: $\alpha = \{0.2, 0.4, 0.6\}$ (**Figure 2**). The time course of the performance—defined as the probability of choosing the more rewarding option $P(Better)$—shows that, for all three learning rates, performance peaks within approximately 10 to 20 trials but more quickly for larger learning rates. However, the overall performance, calculated across the entire block, does not significantly differ among the three learning rates (rank sum test, $p > .05$; inset in **Figure 2b**).

Compared to the performance, the information-theoretic metrics exhibited more distinct trends across the different values of α (**Figure 2c–f**). For example, the entropy of strategy, $H(Str)$, was highest for the RL agent simulated with $\alpha = 0.2$, followed by $\alpha = 0.6$ and $\alpha = 0.4$, which did not differ significantly from each other (inset in **Figure 1c**). Moreover, the trajectory of $H(Str)$, which measures the entropy of stay/switch strategy, decreased over trials within a block and peaked following the reversal—when the better and worse options were swapped. The normalized mutual information between reward and strategy (n-MIRS) revealed that information gained from reward feedback was highest for $\alpha = 0.6$, followed by 0.4, and lowest for 0.2 (**Figure 1d**), consistent with the interpreting the learning rate as a measure of how strongly choices are adjusted based on reward feedback. The normalized mutual information between choice option and strategy (n-MIOS) indicated that the agents gradually learned to identify and persist with the better option over trials, as reflected by the increasing $P(Better)$ and decreasing $H(Str)$ (**Figure 1e**). Finally, the normalized mutual information between option-reward combinations and strategy (n-MIROS) revealed that, despite differences in their temporal trajectories, the overall information shared between these variables plateaued at similar levels across all three learning rates (**Figure 2f**). However, the distinct temporal dynamics led to significant differences among agents as reflected in the average metrics values computed over the entire block (inset in **Figure 2f**).

Overall, this example illustrates that information-theoretic metrics can detect subtle variations in underlying mechanisms—such as variations in the learning rates—that simpler measures like performance may fail to capture (compare the inset in **Figure 2b** with those in **Figure 2c–f**). In the following four sets of simulations, we further show how these information-theoretic metrics can differentiate between the choice behavior of various RL models used as ground truths.

3.1. Revealing Positivity Bias

In the first set of simulations, we examined whether positivity bias—formally defined as a larger learning rate for positive outcomes compared to negative outcome ($\alpha_+ > \alpha_-$)—can be captured using information-theoretic metrics. To that end, we used a standard RL (**Figure 3a**) to simulate choice behavior in a probabilistic reversal learning task under different reward scheudles: 80/20% and 40/10% reward schedules. We explored different combination of α_+ and α_- values to examine how asymmetric learning rates influence behavior across these environments.

We found that that higher positivity bias (larger $\Delta\alpha = \alpha_+ - \alpha_-$) resulted in higher performance across the two reward environments, especially for the 40/10% schedule (Pearson's correlation between $\Delta\alpha$ and $P(Better)$; 40/10%: r = .904, $p = 4.32\times10^{-108}$; 80/20%: r = .702, $p = 2.88\times10^{-44}$) (**Figure 3b**). This is consistent with the previous literature [27,44,45,76], demonstrating that positivity bias can improve performance in terms of reward harvest.

Because learning rates control how value estimates are updated in response to reward feedback, we hypothesized that information-theoretic metrics related to reward—specifically ERDS and n-MIRS—would be sufficient to detect positivity bias in the choice behavior. Following this intuition, we first compared the choice behavior of the standard RL agents simulated with selected values of α_+ and α_- that exemplify "optimistic" ($\alpha_+ > \alpha_-$), "neutral" ($\alpha_+ = \alpha_-$), and "pessimistic" ($\alpha_+ < \alpha_-$) tendencies **Figure 3c**).

We found that the entropy of stay/switch strategy conditioned on reward (ERDS) was highest for the most pessimistic agent (cyan lines in **Figure 3d**), whereas the normalized mutual information between reward and strategy (n-MIRS) was lowest for the same agent (cyan lines in **Figure 3e**). However, metrics computed from the neutral agent (black lines) and the two optimistic agents (red and magenta lines) were less distinguished and plateaued at similar values. When simulating the full grid of α_+ and α_- values, we observed that ERDS was significantly correlated with $\Delta\alpha$, with higher positivity bias associated with lower ERDS (r = -.778, $p = 6.25\times10^{-60}$; **Figure 3f**). In contrast, n-MIRS was only weakly correlated with $\Delta\alpha$ (r = -.137, p = .020), and the direction of this relationship was opposite to that observed in the example shown in **Figure 3e**, suggesting an overall nonlinear relationship. Therefore, while a stronger positivity bias is generally associated with lower ERDS, the absolute values of these metrics alone do not reliably indicate whether a given behavior reflects positivity bias.

To identify candidate metrics that may be more informative of positivity bias, we next examined the decomposition of the above metrics based on reward outcome (reward vs. no reward). Intuitively, there is greater uncertainty about whether the agents will stay/switch following unrewarded trials compared to rewarded ones. Over time, agents learn to maximize reward by adopting a stay-dominated strategy following rewarded trials, as they continue selecting the better option. In contrast, after unrewarded trials, agents are more likely to switch. However, if the chosen option still has a relatively high reward probability (i.e., it is the better option), the likelihood of staying after a no-reward trial can also increase over time.

Consistent with this notion, we found that the trajectory of the decomposition of mutual information metric following no reward (n-MIRS- in **Figure 3g**) became negative over time, suggesting that negative reward outcomes became less informative about the agents' subsequent strategy (i.e., $H(Str) < H(Str|R = loss)$). However, during the early portion of the block, unrewarded trials were initially informative about choice strategy,as indicated by the positive n-MIRS- values. Optimistic and pessimistic agents exhibited distinct temporal dynamics in how quickly this metric shifted toward negative values (shaded gray area in **Figure 3g**). When computed over the first 10 trials, the sign of the n-MIRS– metric was approximately predictive of whether an agent exhibited positivity bias or not (**Figure 3g** inset). Specifically, more optimistic agents exhibited more negative n-MIRS- values, indicating that their tendency to update more strongly after rewarded than unrewarded trials. This makes the no-reward trials less informative of their subsequent strategy. This pattern is consistent with the results shown in **Figure 3b**, where optimistic agents achieved better performance. To confirm this effect across the full parameter space, we computed n-MIRS- as a function of $\Delta\alpha$ and found a strong linear relationship in the 80/20% environment (r = -.926, p = 1.27×10^{-123}), as well as in the other two tested reward environments (60/40% and 40/10%; **Figure 3h**). These results suggest that the decomposed mutual information measures—specifically n-MIRS- computed from the early period of each block—can be used to predict the presence of positivity bias. To test this idea, we conducted a classification analysis in which we randomly sampled groups of optimistic and pessimistic agents and used the information-theoretic metrics to predict whether a given metric profile originated from an optimistic agent using a linear decoder (**Figure 3i**; see Section 2.3.1 for details).

We found that the n-MIRS- metric alone, when computed from the first 10 trials, yielded high cross-validated decoding accuracy across different environments (**Figure 3j**), significantly exceeding chance level (signed-rank test against 0.50; $p < .001$ for all environments). Including all available metrics in the decoder led to a small but significant improvement in the decoding accuracy (signed-rank test; $p = .0011$ for 50/10%, $p < .001$ for all other environments). In comparison, a decoder based on logistic regression weights achieved comparable accuracies (white bars in **Figure 3j**), indicating that the information-theoretic metrics contain comparable information to traditional regression-based measures—while offering more interpretable connections to underlying learning and decision-making processes. Overall, these results illustrate that information-theoretic metrics contain sufficient information to determine the presence of positivity bias when the ground truths are provided.

3.2. Revealing Reward-Dependent Metaplasticity

One potential neural mechanism underlying differential learning rates is reward-dependent metaplasticity, which generates dynamic learning rates by naturally adapting to the recent history of rewards in the environment [46,47,73,77]. To test whether such dynamic changes in the learning rates can be detected using information-theoretic metrics, we implemented a variant of the metaplastic model proposed by Farashahi et al. [46]. In our simulations, we assumed that each of the two choice options is assigned a set of metaplastic synapses which undergo transitions in response to reward feedback (**Figure 4b**; section 2.3.2). By computing the "effective" learning rate—defined as the overall rate of value update on each trial—we found that the metaplastic model exhibited diverging learning rates following rewarded and unrewarded trials, with $\alpha_+ > \alpha_-$ (solid lines in **Figure 4c**), consistent with previous findings [46]. The difference between the effective α_+ and α_- increased over time and plateaued just before the reversal (red solid curve in **Figure 4c**), driven primarily by an increase in α_+ rather than a decrease in α_-. This happens because, as more reward is obtained, the "weak" synapses encoding the better option mostly occupy the most unstable meta-states w_1 (**Figure 4b**), which has the highest transition probability α_1 toward the "strong" state s_1, resulting in a high effective α_+.

Using the choice behavior of this metaplastic agent as ground truth, we fitted a 'plastic' model (standard RL) with differential learning rates to the simulated choice data from each block to test whether the fitted parameters reflected the pattern $\alpha_+ > \alpha_-$ (noting that these learning rates are constant by definition in this model). The results confirmed this pattern (red and blue dashed lines in **Figure 4c**), suggesting that empirically observed trends of $\alpha_+ > \alpha_-$ in previous studies may, in fact, reflect underlying reward-dependent metaplasticity.

To test whether the distinct learning mechanisms of the plastic and metaplastic models can be identified from choice behavior, we simulated the choice behavior of a metaplastic model and two plastic models, and then computed information-theoretic metrics from the simulated behavior. Specifically, we compared metrics across three models: (1) a plastic model with a single $\alpha = 0.3$ (plastic 1-α), (2) a metaplastic model initialized with the same effective learning rate of 0.3 on the first trial, and (3) a plastic model with differential learning rates (plastic 2-α) estimated from the metaplastic agent.

Comparison of the entropy of choice strategy indicated that the metaplastic agent was more consistent in its strategy compared to both plastic agents (rank sum test on $H(Str)$: vs. 1-α: $p = 4.43\times10^{-28}$; vs. 2-α: $p = 2.43\times10^{-24}$; **Figure 4d**). Moreover, the n-MIOS metric revealed that the metaplastic agent had the advantage of increasing the mutual information between the choice of the better option and subsequent strategy (**Figure 4e**), reflected in its superior performance relative to both plastic models (rank sum test on n-MIOS: vs. 1-α: $p = 5.58\times10^{-4}$; vs. 2-α: $p = .00280$). The two plastic agents (1-α and 2-α) were less distinguishable from each other in terms of $H(Str)$ ($p = .0120$), n-MIOS ($p = .683$), and overall performance ($p = .848$).

These results suggest that, although the parameters of the plastic 2-α model were directly estimated from the metaplastic agent, differences in the underlying learning mechanisms still give rise to distinct patterns in the information-theoretic metrics. This highlights the potential utility of such metrics in model recovery and validation—for example, by assessing whether a given candidate RL model can reproduce the information-theoretic metrics observed in the empirical data [29].

Given the observed differences between plastic and metaplastic models, we next quantified the extent to which the information-theoretic metrics discriminate between different models. We hypothesized that these metrics could be used to decode changes in the learning rates corresponding to positivity bias, using an approach similar to that employed in the previous section. To obtain highest accuracy, we utilized all available information-theoretic metrics and their decompositions to train a linear decoder to discriminate between optimistic and pessimistic agents (see section 2.3.2 in *Materials and Methods* for more details). We then applied this decoder to the metrics generated by the three models described above and computed the posterior probability that the decoder would classify the behavior as exhibiting positivity bias.

We found that the posterior probabilities derived from the metaplastic model (magenta line in **Figure 4f**) exhibited temporal dynamics closely resembling those of effective learning rates (**Figure 4c**). Specifically, as trials progressed, the decoder increasingly indicated that the behavior originated from an optimistic agents with $\alpha_+ > \alpha_-$. In contrast, the posterior probabilities for the two plastic agents remained relatively constant across the block. The plastic 1-α agent (black line in **Figure 4f**) maintained performance at chance level, as it was neither optimistic nor pessimistic by design. On the other hand, choice behavior of the plastic 2-α agent (gray line in **Figure 4f**) was classified as being optimistic, consistent with its estimated learning rates reflecting $\alpha_+ > \alpha_-$ (dashed lines in **Figure 4c**).

3.3. Revealing Behavioral Adjustments due to Reward Harvest Rate

Next, we investigated how information-theoretic metrics can be used to detect the influence cumulative overall reward feedback on all options (i.e., reward harvest rate) on an agent's choice strategy. To that end, we used two types of RL model to generate choice behavior: (1) the standard RL with a single learning rate α, and (2) the augmented RL model incorporating a reward harvest rate component (**Figure 5a**). Briefly, the augmented model tracks an average reward rate H, which was not tied to any specific choice option, and uses it to modulate the agent's win-stay and lose-switch tendencies (see section 2.3.3 in *Materials and Methods* for more details). To quantify the effect of reward harvest rate on choice behavior, we defined a discrete variable referred to as the *reward harvest state* (RHS)—a binary indicator of whether the reward rate on a given trial was above or below the block median [35].

As shown for an example block, high or low reward harvest state (RHS) did not consistently coincide with reward (or no reward) on the previous trial (**Figure 5b**), given the weak correlation between the reward rate and binary reward feedback (r = .107, p = .343). Exploiting this property, we used RHS as an additional conditioning variable for the strategy Str, and measured the joint influence of previous reward and RHS on the agent's subsequent strategy using the mutual information metric n-MIRHS (see Section 2.3.3 for more details). The trajectory of this metric revealed that overall information was significantly higher for the agent with the additional reward harvest rate component (*RL with RHR* agent) compared to the standard RL agent (rank sum test, $p = 2.56\times10^{-34}$), reflecting the added influence of RHS on choice strategy **Figure 5c**.

To gain insight on how the high and low RHS states influence behavior, we next examined the decomposition of n-MIRHS based on RHS, corresponding to the influence of high (n-MIRHS$_{high}$, **Figure 5d**) and low RHS (n-MIRHS$_{low}$, **Figure 5e**). We found that the shared information between the *high* state and choice strategy was overall comparable across models but significantly higher for the standard RL agent than for the *RL with RHR* agent (difference in mean n-MIRHS$_{high}$ = 0.0445; rank sum test, $p = 4.67\times10^{-34}$). In contrast, n-MIRHS$_{low}$ revealed a more pronounced distinction: it was significantly positive for the *RL with RHR* agent (signed-rank test, $p = 3.90\times10^{-18}$), but not significantly different from zero for the standard RL model (signed-rank test, p = .064; mean = 0.0019; inset in **Figure 5e**).

These results indicate that during the *high* state—when the reward rate was higher than usual due to successful learning—both models showed positive mutual information between the high RHS and subsequent strategy. In contrast, during the *low* state—when the agents were receiving rewards at a lower-than-usual rate—RHS was not informative about whether the standard RL agent would stay or switch on the next trial. This was not the case for the *RL with RHR* agent, for which n-MIRHS$_{low} > 0$. Further decompositions based on reward feedback revealed that this effect was driven by a larger magnitude of n-MIRHS$_{low+}$ (following rewarded trials) compared to n-MIRHS$_{low-}$ (following unrewarded trials; compare red and blue bars in the **Figure 5e** inset for the RL with RHR agent). This pattern reflects an overall positive shift in information content resulting from the architecture of the RL with RHR model, which modulates stay/switch tendency as a function of the reward harvest rate H.

Lastly, to gain further insight into how the n-MIRHS metric rely on the model parameters, we simulated choice behavior of the model with additional modulation by reward harvest rate using the full range of α and β_H values. Using standardized regression, we then examined the contribution of the β_H parameter, which determines the influence of reward harvest rate on choice behavior ($\beta_H = 0$ corresponds to the standard RL. We found that, consistent with the example in (**Figure 5c**), larger values of β_H predicted overall higher values of n-MIRHS (standardized regression of n-MIRHS on β_H, α, and their interaction: b_1 = 0.366, $p = 3.13\times10^{-216}$). Similar regression analyses on each of the decompositions yielded results consistent with **Figure 5d,e**. More specifically, β_H did not significantly affect the n-MIRHS decomposition for the high reward state (standardized coefficient predicting n-MIRHS$_{high}$: b_1 = 0.0150, p = .329) nor its further reward-based decompositions (n-MIRHS$_{high+}$:

b_1 = 0.024, p = .109; n-MIRHS$_{high-}$: b_1 = -0.00665, p = .674). In contrast, higher values of β_H predicted significantly larger n-MIRHS$_{low}$ (b_1 = 0.464, $p = 4.37\times10^{-308}$). The decompositions based on reward feedback revealed that this effect was strongest for n-MIRHS$_{low-}$, which was significantly higher for larger values of β_H (standardized coefficient: b_1 = 0.523, $p = 1.53\times10^{-253}$; dashed lines in **Figure 5f**, right). In comparison, n-MIRHS$_{low+}$ was also significantly predicted by β_H, but to a lesser degree (b_1 = 0.121, $p = 1.90\times10^{-19}$; solid lines in **Figure 5f**, right). These results suggest that the influence of reward harvest rate on promoting heuristic win-stay/lose-switch is primarily mediated by increased information value of no reward in the low reward state. That is, the larger β_H, corresponding to the effect of reward rate, was associated with less negative n-MIRHS$_{low-}$ toward zero, thus increasing the informativeness of receiving no reward during this state. This can be used to detect modulations of choice behavior by reward harvest rate as these effects can change across blocks of trials.

3.4. Revealing the Presence of Alternative Learning Strategies in Multidimensional Reward Environments

Finally, we examined whether the information-theoretic metrics can be utilized to identify the presence of alternative learning strategies in multidimensional reward environments. In naturalistic settings, reward outcomes often depend on distinct features or attributes of the available choice options. To mimic such scenarios, we simulated a task in which only one of the two choice attributes—either the shape or the color of the stimuli—was associated with reward, thereby introducing uncertainty about which attribute was predictive of reward outcomes (**Figure 6a**; see section 2.3.4 for more details). We considered three RL agents, each exhibiting distinct types of choice strategy: (1) an agent that only learns and chooses based on the color attribute (*Color-only*), (2) an agent that equally weighs the color and shape attributes without arbitration (*No Arbitration*), and (3) an agent which dynamically arbitrates between two attributes based on the reliability of each attribute in predicting reward (*Dynamic Arbitration*; **Figure 6b**).

Examining the arbitration weight ω across different block types revealed that the *Dynamic Arbitration* agent successfully adapted its strategy to the environment: it consistently increased ω during *Color-to-Color* blocks, indicating a growing reliance on color (magenta curve in **Figure 6e**, top), and shifted ω toward a shape-based strategy following reversals in *Color-to-Shape* blocks (magenta curve in **Figure 6f**, top). When reversals occurred within the color attribute (*Color-to-Color* blocks), the performance trajectories ($P(Better)$) were qualitatively similar across all agents, as expected. In each model, performance dropped immediately after reversals followed by gradual recovery (**Figure 6e**, bottom). As expected, the *Color-only* agent reached its peak performance more quickly (**Figure 6e**, bottom), as it did not consider shape values, which were irrelevant in this condition. During *Color-to-Shape* blocks, when the color information was no longer predictive of rewards after reversals, the *Dynamic Arbitration* agent performed the best (**Figure 6f**, bottom), as this agent appropriately identified the correct reward-predictive attribute by dynamically arbitrating between the two feature dimensions (**Figure 6f**, bottom).

Although the learning strategy of the *Color-only* agent can be readily distinguished by its $P(Better)$ trajectory during the *Color-to-Shape* blocks, distinguishing among all three agents is more challenging in *Color-to-Color* blocks based on performance alone. Therefore, we examined whether information-theoretic metrics could be used to distinguish among the three types of learning strategies represented by the RL agents. Although the underlying arbitration weight ω can, in principle, be estimated through model fitting, behavioral metrics offer a more direct window into learning and decision-making strategies. Moreover, these metrics can be computed over a subset of trials and require far simpler computations compared to model fitting, which depends on continuity in the choice data.

To quantify reward-dependent strategies, we computed the conditional entropy of reward-dependent strategies (ERDS) separately for each attribute—ERDS$_{Shape}$ and ERDS$_{Color}$ (**Figure 7a,b**). We found that during *Color-to-Color* blocks, the entropy of the shape-based strategy (ERDS$_{Shape}$) was close to its maximum value of 1 for all the three agents, reflecting the fact that shape carried no information about reward in this block type (dashed lines in **Figure 7a**). In contrast, the entropy of the color-based strategy (ERDS$_{Color}$) exhibited modulation across trials, with distinct dynamics for

doi:10.20944/preprints202509.0001.v1

each agent (solid lines in **Figure 7a**). Consistent with the performance trajectories, the *Color-only* agent showed the lowest entropy, followed by the *Dynamic Arbitration* agent (rank sum test on ERDS_{Color}, $p = 7.98 \times 10^{-34}$). To assess the relative dominance of the two learning strategies, we computed the difference between the two entropy values, $\mathrm{ERDS}_{Shape} - \mathrm{ERDS}_{Color}$. The resulting positive values indicated that all three agents correctly prioritized the color-based learning strategy during this block type, as reflected by lower entropy for the relevant dimension (i.e., color) **Figure 7c**). These two metrics thus provide insight beyond the trajectories of arbitration weight ω (**Figure 6e** top), which does not directly capture changes in the learned values that could drive choice behavior.

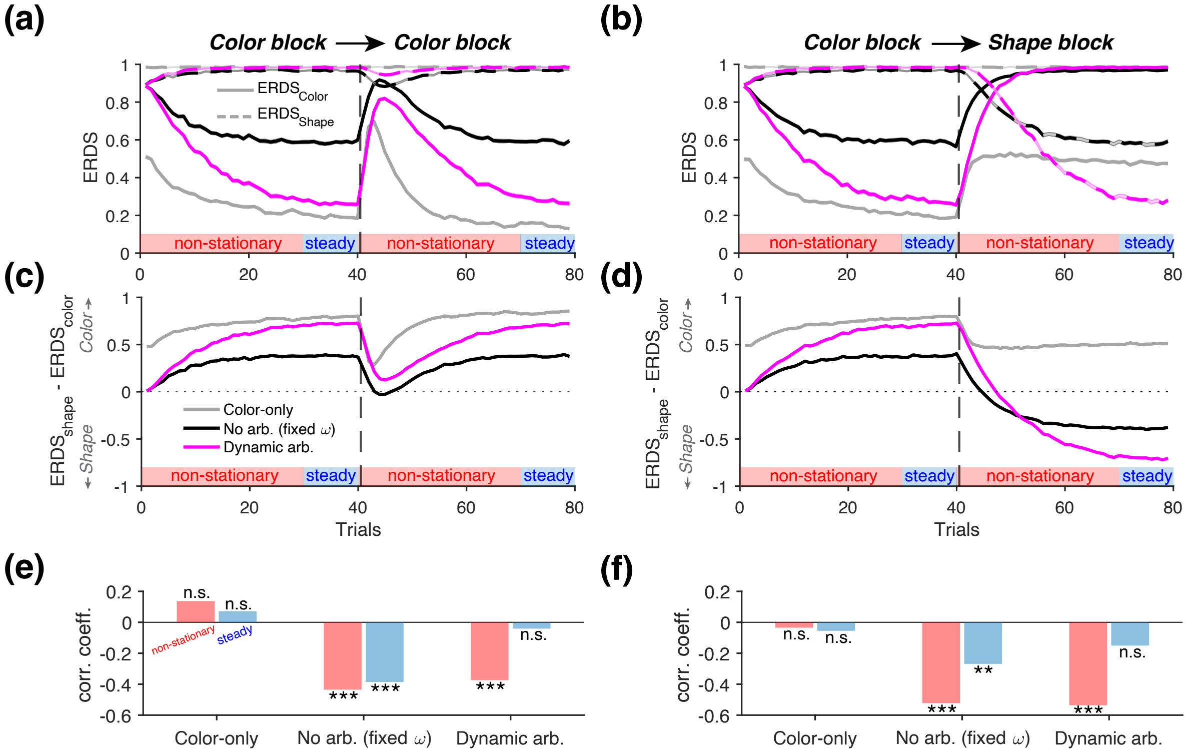

Figure 7. Distinct predictions of information-theoretic metrics for adopting and arbitrating between alternative strategies in a multidimensional reward environment. (**a**) Time courses of the conditional entropy of reward-dependent strategy based on color ($\mathrm{ERDS}_{Color} = H(Str_{Color}|R)$, solid lines) and shape ($\mathrm{ERDS}_{Shape} = H(Str_{Shape}|R)$, dashed lines) during the Color-to-Color blocks. Line colors indicate the three simulated agents (gray: *Color-only*; black: *No Arbitration*; magenta: *Dynamic Arbitration*). (**b**) Similar to (a) but for the Color-to-Shape (feature reversal) blocks. (**c**) Time course of the difference $\mathrm{ERDS}_{Shape} - \mathrm{ERDS}_{Color}$ during the Color-to-Color blocks, measuring the relative dominance of two learning strategies. Higher values indicate dominance of the color-based strategy. (**d**) Similar to (c) but for the Color-to-Shape (feature reversal) blocks. (**e–f**) Correlation coefficients between ERDS_{Color} and ERDS_{Shape} during the Color-to-Color (e) and the Color-to-Shape (feature reversal) blocks (f), computed separately from the "non-stationary" (first 30 trials of each block) or "steady" portion of the block (last 10 trials before the reversal or end of the block). Asterisks indicate significance level (*: $p < .05$; **: $p < .01$; ***: $p < .001$; *n.s.*: not significant). The *Color-only* agent showed no interaction between the two strategies, with no significant correlations observed between ERDS_{Color} and ERDS_{Shape} in any block or phase. The *No Arbitration* agent (with fixed $\omega = 0.5$) exhibited strong negative interaction between the two strategies, as reflected by significant negative correlations during both steady and non-stationary portions of the blocks. The *Dynamic Arbitration* agent showed significant negative interaction during the non-stationary phase, but this correlation became non-significant during the steady phase as one strategy began to dominate.

During the *Color-to-Shape* blocks, the entropy values (ERDS_{Shape}) indicated near-random use of the shape-based strategy for all three agents prior to the reversal (dashed lines in **Figure 7b**). After the reversal—when shape attribute became informative of reward—the *Dynamic Arbitration* showed the most pronounced adjustment, characterized by a decreased reliance on the color-based strategy (reflected by as drop in ERDS_{Color}; solid magenta line in **Figure 7b**), and an increased reliance on

doi:10.20944/preprints202509.0001.v1

the shape-based strategy (reflected by a rise in ERDS$_{Shape}$; dashed magenta line). The difference in the two entropy values further revealed that only the agents capable of learning both attributes (*No Arbitration* and *Dynamic Arbitration*) were able to shift their strategies after the reversal, as indicated by ERDS$_{Color}$ $>$ ERDS$_{Shape}$ (**Figure 7d**). In contrast, the *Color-only* agent exhibited ERDS$_{Color}$ $<$ ERDS$_{Shape}$ throughout the entire block.

To quantify potential interactions between the color-based and shape-based learning strategies, we measured the correlation between ERDS$_{Shape}$ and ERDS$_{Color}$, computed separately from two distinct periods within each block. "Non-stationary" trials were defined as the first 30 trials at the beginning of each block and those immediately following the reversal (highlighted by red bars in **Figure 7a,b**). In contrast, "steady" trials corresponded to the final 10 trials before the reversal or at end of each block, during which performance had plateaued (blue bars in **Figure 7a–d**).

This analysis revealed distinct patterns of interaction between the two learning strategies across the three agents (**Figure 7e,f**). More specifically, the *Color-only* agent showed no significant correlation between ERDS$_{Shape}$ and ERDS$_{Color}$ during either periods of the *Color-to-Shape* blocks (non-stationary: r = .137, p = .174; steady: r = .072, p = .477; **Figure 7e**). This result is expected, as the *Color-only* agent does not employ any shape-based strategy, which renders response to reward based on shape effectively random. In contrast, the *No Arbitration* agent exhibited competitive interaction between the two strategies, as indicated by a significant negative correlation between ERDS$_{Shape}$ and ERDS$_{Color}$ (non-stationary: r = -.435, p = 6.08×10^{-6}; steady: r = -.386, p = 7.43×10^{-5}; **Figure 7e**). This pattern reflects the agent's decision-making strategy , in which color and shape information are always weighted equally, regardless of which attribute is currently predictive of reward (fixed $\omega = 0.5$). Interestingly, the *Dynamic Arbitration* agent exhibited different trend of interaction across the two periods of the block. During the non-stationary phase, this model exhibited a significant negative correlation between ERDS$_{Shape}$ and ERDS$_{Color}$ (r = -.374, p = 1.28×10^{-4}). However, this correlation disappeared during the steady phase of the block (r = -.0409, p = .686), when the color-based strategy became dominant (i.e., $\omega > 0.5$). This shift in correlation pattern can be used to detect dynamic arbitration between alternative learning strategies.

We found overall consistent results during the *Color-to-Shape* blocks (**Figure 7f**). The *Color-only* agent showed no significant interaction between strategies in either period (non-stationary: r = -.0346, p = .733; steady: r = -.055, p = .587). In contrast, the *No Arbitration* agent exhibited significant negative correlations during both phases of the block (r = -.522, p = 2.59×10^{-8}; steady: r = -.269, p = .00685). Notably, only the *Dynamic Arbitration* agent—capable of adapting its behavior based on relative reliability of the two attributes in predicting reward—exhibited a shift in the interaction pattern over time. Specifically, it showed a significant negative correlation during the non-stationary period (r = -.536, p = 9.19×10^{-9}), which weakened and became non-significant during the steady phase (r = -.151, p = .135), consistent with the pattern observed in the *Color-to-Color* blocks. Together, these results demonstrate that a negative correlation between ERDS$_{Shape}$ and ERDS$_{Color}$ during the non-stationary phase and the absence of such a correlation during the steady phase is indicative of dynamic arbitration between alternative learning strategies.

4. Discussion

In this study, we demonstrate how behavioral metrics inspired by information theory can be used to identify certain learning and decision-making mechanisms. In particular, we applied metrics based on conditional entropy, normalized mutual information, and their decompositions based on specific outcomes to choice behavior during two variants of the probabilistic reversal learning task—a widely adopted paradigm for studying cognitive flexibility across species. Using these metrics, we investigated whether specific neural or computational mechanisms—specified by the reinforcement learning (RL) models serving as ground truth—could be inferred from choice behavior. To that end, we examined positivity bias, gradual changes in the learning rates due to reward-dependent metaplasticity, the influence of reward harvest rate, and the adoption and arbitration between alternative learning strategies in multidimensional environments.

One of the key strengths of the proposed information-theoretic metrics lies in their versatility and flexibility. In contrast, fitting RL models to choice behavior in order to identify underlying mechanisms requires a continuous stream of choice and reward data, with the precision of the estimated parameters heavily dependent on the amount of available data. As our results demonstrate, information-theoretic metrics circumvent these limitations, as they can be computed from as few as several dozen trials, even when drawn from non-contiguous segments of the data. Directly comparing the information content of our metrics with regression weights for predicting choice (**Figure 3j**), we found similar decoding performance from both types of measures. However, information-theoretic metrics offer the distinct advantage of being simple and model-free, whereas regression-based features are inherently model-dependent——requiring decisions about which predictors to include, how many lags to consider, how to evaluate model fit and statistical significance, and how to interpret the resulting regression weights.

Moreover, time courses of information-theoretic metrics can be computed for each trial point by concatenating trial-wise data across different blocks. This approach serves as a useful visualization tool for examining behavioral changes over time. Finally, the analysis in **Figure 7** demonstrates how metrics computed from different periods within a block can provide insight into temporal changes in learning and choice strategies. Taken together, the efficiency and flexibility of these metrics make them valuable model-free tools for identifying and testing hypotheses about the mechanisms underlying learning and decision making.

Another important advantage of information-theoretic metrics is their adaptability to target specific variables of interest within a given study. As illustrated in simulations of reward harvest rate effects on behavior (**Figure 5**), the key variable—reward harvest state—can be incorporated into the mutual information metric n-MIRHS to assess how reward harvest rate influences choice behavior. This measure revealed the role of reward harvest rate in shaping choice strategy under varying reward contingencies by quantifying the information flow from prior reward outcomes and reward states into the agent's choice strategy. Therefore, these metrics can be flexibly formulated to match the demands of a given research context and provide quantifiable insights into the role of specific variables.

In turn, the insights provided by these metrics can guide the development of improved models that better capture key aspects of behavioral data [29]. As numerous studies have pointed out [78–80], the best-fitting model among a set of candidates may still fail to reproduce important features of choice behavior. Therefore, validating candidate models through posterior predictive checks remains a critical step in modeling [26,81]. Information-theoretic metrics offer ideal tools for this process, as they provide summary statistics specific to the variables of interest. For example, different RL models could be validated by comparing information-theoretic metrics computed from simulated data with those derived from empirical data [29,38]. By quantifying the divergence between empirical observations and model-generated data, this comparison both reveals a model's limitations and pinpoints the mechanisms that must be added for improvement [29]. Such an approach is especially valuable when standard model-comparison metrics are inconclusive—for example, with small or noisy datasets that inherently favor simpler, low-parameter models. In those cases, one can use information-theoretic measures to constrain the candidate models by selecting the one whose simulated statistics most closely match the empirical data. Lastly, although this study focuses on RL models, the same framework can be applied to other model classes—such as those based on Bayesian inference [32,82].

In addition to serving as useful tools for model discovery and validation, information-theoretic metrics may also contribute to data-driven approaches in studies involving choice behavior. Our decoding analyses in Sections 3.1 and 3.2 serves as a proof-of-concept, showing that the information-theoretic measures contain sufficient information to detect positivity bias when the ground truths are known. This example highlights the broader potential of the metrics in detecting latent cognitive states or mechanisms other than positivity bias (which may lack ground truths in real data). For example, one could train a decoder to predict experimentally defined task states from either behavioral measures (reaction time, reward rate, pupil dilation) or neural recordings (spike trains, EEG, fMRI) by using information-theoretic features extracted from choice data. In this approach, the experimentally defined states serve as the ground-truth labels during training, and the decoder's ability to recover those labels is then evaluated on held-out behavioral (or neural) data. Future research could apply this model-agnostic yet interpretable framework to infer latent cognitive or neural states.

There are some challenges and considerable potential for extending the information-theoretic framework presented here. For example, while our entropy measures quantify the overall consistency in the stay/switch strategy in response to reward and other relevant variables, they do not directly capture directionality—that is, they do not indicate what causes what. As a result, these metrics should be considered alongside other measures to provide a more complete picture. This limitation makes interpreting mutual information versus conditional entropy particularly challenging. Moreover, while information-theoretic metrics capture sensitivity to reward feedback, they cannot be uniquely mapped onto the learning rates. Finally, although we have focused on conditional entropy and normalized mutual information, other behavioral metrics based on other concepts from information theory can be developed including transfer entropy [83,84], mutual information between discrete and continuous variables [85–87], and partial information decomposition [88,89], among others. Incorporating these extensions would allow the behavioral metrics to be generalized to more complex tasks, including those involving more than two alternative options, higher-dimensional feature spaces, and task contexts.

Author Contributions: Conceptualization, J.H.W. and A.S.; methodology, J.H.W. and A.S.; software, J.H.W. and L.B.; validation, J.H.W.; formal analysis, J.H.W., L.B., and A.S.; investigation, J.H.W. and L.B.; resources, J.H.W. and L.B.; data curation, J.H.W.; writing—original draft preparation, J.H.W., L.B., and A.S.; writing—review and editing, J.H.W., L.B., and A.S.; visualization, J.H.W.; supervision, A.S.; project administration, J.H.W. and A.S.; funding acquisition, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institutes of Health grant R01 DA047870 to A.S.

Data Availability Statement: The analysis codes used for this study is available at https://github.com/DartmouthCCNL/woo_etal_entropy.

Acknowledgments: We thank the members of the Computational Cognitive Neuroscience Lab, Peter Duggins, Michael Chong Wang, Aryan Yazdanpanah, and Atlas Shahamati, for their helpful comments on the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lungarella, M.; Sporns, O. Mapping information flow in sensorimotor networks. *PLoS computational biology* **2006**, *2*, e144.
2. Palmigiano, A.; Geisel, T.; Wolf, F.; Battaglia, D. Flexible information routing by transient synchrony. *Nature Neuroscience* **2017**, *20*, 1014–1022. Publisher: Nature Publishing Group, https://doi.org/10.1038/nn.4569.
3. Honey, C.J.; Kötter, R.; Breakspear, M.; Sporns, O. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc Natl Acad Sci U S A* **2007**, *104*, 10240–10245. https://doi.org/10.1073/pnas.0701519104.
4. Stramaglia, S.; Wu, G.R.; Pellicoro, M.; Marinazzo, D. Expanding the transfer entropy to identify information circuits in complex systems. *Phys. Rev. E* **2012**, *86*, 066211. Publisher: American Physical Society, https://doi.org/10.1103/PhysRevE.86.066211.
5. Novelli, L.; Wollstadt, P.; Mediano, P.; Wibral, M.; Lizier, J.T. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Network Neuroscience* **2019**, *3*, 827–847. https://doi.org/10.1162/netn_a_00092.
6. Vicente, R.; Wibral, M.; Lindner, M.; Pipa, G. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience* **2011**, *30*, 45–67.
7. Ursino, M.; Ricci, G.; Magosso, E. Transfer Entropy as a Measure of Brain Connectivity: A Critical Analysis With the Help of Neural Mass Models. *Front Comput Neurosci* **2020**, *14*, 45. https://doi.org/10.3389/fncom.2020.00045.

8. Strong, S.P.; Koberle, R.; De Ruyter Van Steveninck, R.R.; Bialek, W. Entropy and Information in Neural Spike Trains. *Phys. Rev. Lett.* **1998**, *80*, 197–200. https://doi.org/10.1103/PhysRevLett.80.197.
9. Gourévitch, B.; Eggermont, J.J. Evaluating Information Transfer Between Auditory Cortical Neurons. *Journal of Neurophysiology* **2007**, *97*, 2533–2543. Publisher: American Physiological Society, https://doi.org/10.1152/jn.01106.2006.
10. Cofré, R.; Maldonado, C. Information Entropy Production of Maximum Entropy Markov Chains from Spike Trains. *Entropy (Basel)* **2018**, *20*, 34. https://doi.org/10.3390/e20010034.
11. Shorten, D.P.; Spinney, R.E.; Lizier, J.T. Estimating Transfer Entropy in Continuous Time Between Neural Spike Trains or Other Event-Based Data. *PLOS Computational Biology* **2021**, *17*, e1008054. Publisher: Public Library of Science, https://doi.org/10.1371/journal.pcbi.1008054.
12. Strange, B.A.; Duggins, A.; Penny, W.; Dolan, R.J.; Friston, K.J. Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Networks* **2005**, *18*, 225–230. https://doi.org/10.1016/j.neunet.2004.12.004.
13. Bach, D.R.; Dolan, R.J. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat Rev Neurosci* **2012**, *13*, 572–586. Publisher: Nature Publishing Group, https://doi.org/10.1038/nrn3289.
14. Sayood, K. Information Theory and Cognition: A Review. *Entropy (Basel)* **2018**, *20*, 706. https://doi.org/10.3390/e20090706.
15. Seiler, J.P.H.; Dan, O.; Tüscher, O.; Loewenstein, Y.; Rumpel, S. Experienced entropy drives choice behavior in a boring decision-making task. *Sci Rep* **2022**, *12*, 3162. Publisher: Nature Publishing Group, https://doi.org/10.1038/s41598-022-06861-w.
16. Jepma, M.; Nieuwenhuis, S. Pupil diameter predicts changes in the exploration-exploitation trade-off: evidence for the adaptive gain theory. *J Cogn Neurosci* **2011**, *23*, 1587–1596. https://doi.org/10.1162/jocn.2010.21548.
17. Wang, M.Z.; Hayden, B.Y. Monkeys are curious about counterfactual outcomes. *Cognition* **2019**, *189*, 1–10. https://doi.org/10.1016/j.cognition.2019.03.009.
18. Woo, J.H.; Azab, H.; Jahn, A.; Hayden, B.; Brown, J.W. The PRO model accounts for the anterior cingulate cortex role in risky decision-making and monitoring. *Cognitive, Affective, & Behavioral Neuroscience* **2022**, *22*, 952–968. Number: 5 Publisher: Springer, https://doi.org/10.3758/s13415-022-00992-3.
19. Lee, D.; Conroy, M.L.; McGreevy, B.P.; Barraclough, D.J. Reinforcement learning and decision making in monkeys during a competitive game. *Cognitive Brain Research* **2004**, *22*, 45–58. https://doi.org/10.1016/j.cogbrainres.2004.07.007.
20. Lee, D.; McGreevy, B.P.; Barraclough, D.J. Learning and decision making in monkeys during a rock–paper–scissors game. *Cognitive Brain Research* **2005**, *25*, 416–430. https://doi.org/10.1016/j.cogbrainres.2005.07.003.
21. Takahashi, H.; Izuma, K.; Matsumoto, M.; Matsumoto, K.; Omori, T. The Anterior Insula Tracks Behavioral Entropy during an Interpersonal Competitive Game. *PLoS One* **2015**, *10*, e0123329. https://doi.org/10.1371/journal.pone.0123329.
22. Sutton, R.S.; Barto, A.G. *Reinforcement Learning, second edition: An Introduction*; MIT Press, 2018. Google-Books-ID: uWV0DwAAQBAJ.
23. Dayan, P.; Daw, N.D. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience* **2008**, *8*, 429–453.
24. Niv, Y. Reinforcement learning in the brain. *Journal of Mathematical Psychology* **2009**, *53*, 139–154. https://doi.org/10.1016/j.jmp.2008.12.005.
25. Jensen, K.T. An introduction to reinforcement learning for neuroscience, 2024. arXiv:2311.07315 [q-bio], https://doi.org/10.48550/arXiv.2311.07315.
26. Eckstein, M.K.; Wilbrecht, L.; Collins, A.G. What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences* **2021**, *41*, 128–137. https://doi.org/10.1016/j.cobeha.2021.06.004.
27. Nussenbaum, K.; Hartley, C.A. Reinforcement learning across development: What insights can we draw from a decade of research? *Dev Cogn Neurosci* **2019**, *40*, 100733. https://doi.org/10.1016/j.dcn.2019.100733.
28. Bartolo, R.; Averbeck, B.B. Prefrontal cortex predicts state switches during reversal learning. *Neuron* **2020**, *106*, 1044–1054.
29. Trepka, E.; Spitmaan, M.; Bari, B.A.; Costa, V.D.; Cohen, J.Y.; Soltani, A. Entropy-based metrics for predicting choice behavior based on local response to reward. *Nat Commun* **2021**, *12*, 6567. Publisher: Nature Publishing Group, https://doi.org/10.1038/s41467-021-26784-w.

30. Grossman, C.D.; Bari, B.A.; Cohen, J.Y. Serotonin neurons modulate learning rate through uncertainty. *Curr Biol* **2022**, *32*, 586–599.e7. https://doi.org/10.1016/j.cub.2021.12.006.
31. Yang, M.A.; Jung, M.W.; Lee, S.W. Striatal arbitration between choice strategies guides few-shot adaptation. *Nature Communications* **2025**, *16*, 1811. Publisher: Nature Publishing Group, https://doi.org/10.1038/s41467-025-57049-5.
32. Nassar, M.R.; Wilson, R.C.; Heasly, B.; Gold, J.I. An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *J. Neurosci.* **2010**, *30*, 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010.
33. Farashahi, S.; Xu, J.; Wu, S.W.; Soltani, A. Learning arbitrary stimulus-reward associations for naturalistic stimuli involves transition from learning about features to learning about objects. *Cognition* **2020**, *205*, 104425. https://doi.org/10.1016/j.cognition.2020.104425.
34. Cinotti, F.; Coutureau, E.; Khamassi, M.; Marchand, A.R.; Girard, B. Regulation of reinforcement learning parameters captures long-term changes in rat behaviour. *European Journal of Neuroscience* **2024**, *60*, 4469–4490. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.16449, https://doi.org/10.1111/ejn.16449.
35. Wittmann, M.K.; Fouragnan, E.; Folloni, D.; Klein-Flügge, M.C.; Chau, B.K.; Khamassi, M.; Rushworth, M.F. Global reward state affects learning and activity in raphe nucleus and anterior insula in monkeys. *Nature communications* **2020**, *11*, 3771.
36. Lee, S.W.; Shimojo, S.; O'doherty, J.P. Neural computations underlying arbitration between model-based and model-free learning. *Neuron* **2014**, *81*, 687–699.
37. Charpentier, C.J.; Iigaya, K.; O'Doherty, J.P. A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning. *Neuron* **2020**, *106*, 687–699.
38. Woo, J.H.; Costa, V.D.; Taswell, C.A.; Rothenhoefer, K.M.; Averbeck, B.B.; Soltani, A. Contribution of amygdala to dynamic model arbitration under uncertainty. *bioRxiv* **2024**, p. 2024.09.13.612869. https://doi.org/10.1101/2024.09.13.612869.
39. Philippe, R.; Janet, R.; Khalvati, K.; Rao, R.P.; Lee, D.; Dreher, J.C. Neurocomputational mechanisms involved in adaptation to fluctuating intentions of others. *Nature communications* **2024**, *15*, 3189.
40. Collins, A.G.E.; Shenhav, A. Advances in modeling learning and decision-making in neuroscience. *Neuropsychopharmacol.* **2022**, *47*, 104–118. https://doi.org/10.1038/s41386-021-01126-y.
41. Woo, J.H.; Aguirre, C.G.; Bari, B.A.; Tsutsui, K.I.; Grabenhorst, F.; Cohen, J.Y.; Schultz, W.; Izquierdo, A.; Soltani, A. Mechanisms of adjustments to different types of uncertainty in the reward environment across mice and monkeys. *Cogn Affect Behav Neurosci* **2023**, *23*, 600–619. https://doi.org/10.3758/s13415-022-01059-z.
42. Lefebvre, G.; Lebreton, M.; Meyniel, F.; Bourgeois-Gironde, S.; Palminteri, S. Behavioural and neural characterization of optimistic reinforcement learning. *Nat Hum Behav* **2017**, *1*, 1–9. Publisher: Nature Publishing Group, https://doi.org/10.1038/s41562-017-0067.
43. Palminteri, S.; Lefebvre, G.; Kilford, E.J.; Blakemore, S.J. Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology* **2017**, *13*, e1005684.
44. Lefebvre, G.; Summerfield, C.; Bogacz, R. A normative account of confirmation bias during reinforcement learning. *Neural Comput* **2022**, *34*, 307–337. https://doi.org/10.1162/neco_a_01455.
45. Palminteri, S.; Lebreton, M. The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences* **2022**, *26*, 607–621. Publisher: Elsevier, https://doi.org/10.1016/j.tics.2022.04.005.
46. Farashahi, S.; Donahue, C.H.; Khorsand, P.; Seo, H.; Lee, D.; Soltani, A. Metaplasticity as a Neural Substrate for Adaptive Learning and Choice under Uncertainty. *Neuron* **2017**, *94*, 401–414.e6. https://doi.org/10.1016/j.neuron.2017.03.044.
47. Khorsand, P.; Soltani, A. Optimal structure of metaplasticity for adaptive learning. *PLoS Comput Biol* **2017**, *13*, e1005630. https://doi.org/10.1371/journal.pcbi.1005630.
48. Niv, Y.; Daw, N.D.; Joel, D.; Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* **2007**, *191*, 507–520.
49. Sukumar, S.; Shadmehr, R.; Ahmed, A.A. Effects of reward and effort history on decision making and movement vigor during foraging. *Journal of neurophysiology* **2024**, *131*, 638–651.
50. Soltani, A.; Koechlin, E. Computational models of adaptive behavior and prefrontal cortex. *Neuropsychopharmacology* **2022**, *47*, 58–71.
51. O'Doherty, J.P.; Lee, S.W.; Tadayonnejad, R.; Cockburn, J.; Iigaya, K.; Charpentier, C.J. Why and how the brain weights contributions from a mixture of experts. *Neuroscience & Biobehavioral Reviews* **2021**, *123*, 14–23.

52. Leong, Y.C.; Radulescu, A.; Daniel, R.; DeWoskin, V.; Niv, Y. Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron* **2017**, *93*, 451–463.
53. Farashahi, S.; Rowe, K.; Aslami, Z.; Lee, D.; Soltani, A. Feature-based learning improves adaptability without compromising precision. *Nature Communications* **2017**, *8*, 1768. Publisher: Nature Publishing Group, https://doi.org/10.1038/s41467-017-01874-w.
54. Radulescu, A.; Niv, Y.; Ballard, I. Holistic reinforcement learning: the role of structure and attention. *Trends in cognitive sciences* **2019**, *23*, 278–292.
55. Farashahi, S.; Soltani, A. Computational mechanisms of distributed value representations and mixed learning strategies. *Nature Communications* **2021**, *12*, 7191. Publisher: Nature Publishing Group, https://doi.org/10.1038/s41467-021-27413-2.
56. Wang, M.C.; Soltani, A. Contributions of Attention to Learning in Multidimensional Reward Environments **2025**. Publisher: Society for Neuroscience Section: Research Articles, https://doi.org/10.1523/JNEUROSCI.2300-23.2024.
57. Wise, T.; Emery, K.; Radulescu, A. Naturalistic reinforcement learning. *Trends in Cognitive Sciences* **2024**, *28*, 144–158.
58. Yazdanpanah, A.; Wang, M.C.; Trepka, E.; Benz, M.; Soltani, A. Contributions of statistical learning to learning from reward feedback. *bioRxiv* **2024**. Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2024/07/26/2024.04.27.591445.full.pdf, https://doi.org/10.1101/2024.04.27.591445.
59. Izquierdo, A.; Brigman, J.L.; Radke, A.K.; Rudebeck, P.H.; Holmes, A. The neural basis of reversal learning: An updated perspective. *Neuroscience* **2017**, *345*, 12–26. https://doi.org/10.1016/j.neuroscience.2016.03.021.
60. 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* **2016**, *92*, 505–517.
61. Bari, B.A.; Grossman, C.D.; Lubin, E.E.; Rajagopalan, A.E.; Cressy, J.I.; Cohen, J.Y. Stable representations of decision variables for flexible behavior. *Neuron* **2019**, *103*, 922–933.
62. Hamid, A.A.; Pettibone, J.R.; Mabrouk, O.S.; Hetrick, V.L.; Schmidt, R.; Vander Weele, C.M.; Kennedy, R.T.; Aragona, B.J.; Berke, J.D. Mesolimbic dopamine signals the value of work. *Nature Neuroscience* **2016**, *19*, 117–126. Publisher: Nature Publishing Group, https://doi.org/10.1038/nn.4173.
63. Aguirre, C.G.; Woo, J.H.; Romero-Sosa, J.L.; Rivera, Z.M.; Tejada, A.N.; Munier, J.J.; Perez, J.; Goldfarb, M.; Das, K.; Gomez, M.; et al. Dissociable Contributions of Basolateral Amygdala and Ventrolateral Orbitofrontal Cortex to Flexible Learning Under Uncertainty **2024**. Publisher: Society for Neuroscience Section: Research Articles, https://doi.org/10.1523/JNEUROSCI.0622-23.2023.
64. Rutledge, R.B.; Lazzaro, S.C.; Lau, B.; Myers, C.E.; Gluck, M.A.; Glimcher, P.W. Dopaminergic Drugs Modulate Learning Rates and Perseveration in Parkinson's Patients in a Dynamic Foraging Task. *J. Neurosci.* **2009**, *29*, 15104–15114.
65. Grunwald, P.; Vitanyi, P. Shannon Information and Kolmogorov Complexity, 2004. arXiv:cs/0410002, https://doi.org/10.48550/arXiv.cs/0410002.
66. Lizier, J.T. Measuring the Dynamics of Information Processing on a Local Scale in Time and Space. In *Directed Information Measures in Neuroscience*; Wibral, M.; Vicente, R.; Lizier, J.T., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2014; pp. 161–193. Series Title: Understanding Complex Systems, https://doi.org/10.1007/978-3-642-54474-3_7.
67. Yao, Y. Information-theoretic measures for knowledge discovery and data mining. *Entropy measures, maximum entropy principle and emerging applications* **2003**, pp. 115–136.
68. Kvålseth, T.O. On normalized mutual information: measure derivations and properties. *Entropy* **2017**, *19*, 631.
69. Amelio, A.; Pizzuti, C. Correction for Closeness: Adjusting Normalized Mutual Information Measure for Clustering Comparison. *Computational Intelligence* **2017**, *33*, 579–601. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/coin.12100, https://doi.org/10.1111/coin.12100.
70. Todorov, E. Efficient computation of optimal actions. *Proceedings of the national academy of sciences* **2009**, *106*, 11478–11483.
71. Ramírez-Ruiz, J.; Grytskyy, D.; Mastrogiuseppe, C.; Habib, Y.; Moreno-Bote, R. Complex behavior from intrinsic motivation to occupy future action-state path space. *Nature Communications* **2024**, *15*, 6368.
72. Donahue, C.H.; Lee, D. Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. *Nature neuroscience* **2015**, *18*, 295–301.

73. Soltani, A.; Wang, X.J. A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *Journal of Neuroscience* **2006**, *26*, 3731–3744.
74. Soltani, A.; Wang, X.J. Synaptic Computation Underlying Probabilistic Inference. *Nat Neurosci* **2010**, *13*, 112–119. https://doi.org/10.1038/nn.2450.
75. Farashahi, S.; Donahue, C.H.; Hayden, B.Y.; Lee, D.; Soltani, A. Flexible combination of reward information across primates. *Nature human behaviour* **2019**, *3*, 1215–1224.
76. Cazé, R.D.; van der Meer, M.A. Adaptive properties of differential learning rates for positive and negative outcomes. *Biological cybernetics* **2013**, *107*, 711–719.
77. Iigaya, K. Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system. *eLife* **2016**, *5*, e18073. Publisher: eLife Sciences Publications, Ltd, https://doi.org/10.7554/eLife.18073.
78. Nassar, M.R.; Frank, M.J. Taming the beast: extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences* **2016**, *11*, 49–54. https://doi.org/10.1016/j.cobeha.2016.04.003.
79. Palminteri, S.; Wyart, V.; Koechlin, E. The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences* **2017**, *21*, 425–433.
80. Sifar, A.; Srivastava, N. Over-precise Predictions Cannot Identify Good Choice Models. *Computational Brain & Behavior* **2022**, *5*, 378–396. Number: 3 Publisher: Springer, https://doi.org/10.1007/s42113-022-00146-1.
81. Wilson, R.C.; Collins, A.G. Ten simple rules for the computational modeling of behavioral data. *Elife* **2019**, *8*, e49547.
82. Meyniel, F.; Dehaene, S. Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences* **2017**, *114*, E3859–E3868.
83. Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464. https://doi.org/10.1103/PhysRevLett.85.461.
84. Bossomaier, T.; Barnett, L.; Harré, M.; Lizier, J.T. Transfer Entropy. In *An Introduction to Transfer Entropy: Information Flow in Complex Systems*; Bossomaier, T.; Barnett, L.; Harré, M.; Lizier, J.T., Eds.; Springer International Publishing: Cham, 2016; pp. 65–95. https://doi.org/10.1007/978-3-319-43222-9_4.
85. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. https://doi.org/10.1103/PhysRevE.69.066138.
86. Gao, W.; Kannan, S.; Oh, S.; Viswanath, P. Estimating Mutual Information for Discrete-Continuous Mixtures. In Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017, Vol. 30.
87. Ross, B.C. Mutual Information between Discrete and Continuous Data Sets. *PLOS ONE* **2014**, *9*, e87357. Publisher: Public Library of Science, https://doi.org/10.1371/journal.pone.0087357.
88. Williams, P.L.; Beer, R.D. Nonnegative Decomposition of Multivariate Information, 2010. arXiv:1004.2515 [cs], https://doi.org/10.48550/arXiv.1004.2515.
89. Luppi, A.I.; Rosas, F.E.; Mediano, P.A.M.; Menon, D.K.; Stamatakis, E.A. Information decomposition and the informational architecture of the brain. *Trends in Cognitive Sciences* **2024**, *28*, 352–368. Publisher: Elsevier, https://doi.org/10.1016/j.tics.2023.11.005.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.