Neuroscience

Individuals with anxiety and depression use atypical decision strategies in an uncertain world

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint version 1
February 2, 2024 (this version)

Sent for peer review
November 10, 2023

Posted to preprint server
October 24, 2023

Zeming Fang, Meihua Zhao, Ting Xu, Yuhang Li, Hanbo Xie, Peng Quan, Haiyang Geng, Ru-Yuan Zhang

Shanghai Mental Health, School of Medicine, Shanghai Jiao Tong University, Shanghai, 200030, China • Institute of Psychology and Behavioral Science, Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, 200030, China • School of Psychology, South China Normal University, Guangzhou, 510631, China • The Center of Psychosomatic Medicine, Sichuan Provincial Center for Mental Health, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, 611731, China • Centre of Centre for Cognitive and Brain Sciences, Institute of Collaborative Innovation, University of Macau, Macau, 999078, China • Department of Psychology, University of Arizona, Tucson, Arizona, USA • Research Center for Quality of Life and Applied Psychology, Guangdong Medical University, Dongguan, China • Tianqiao and Chrissy Chen Institute for Translational Research, Shanghai, 200040, China • Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, School of Psychology and Cognitive Science, East China Normal University, Shanghai, 200241, China

Abstract

The theory of optimal learning proposes that an agent should increase or decrease the learning rate in environments where reward conditions are relatively volatile or stable, respectively. Deficits in such flexible learning rate adjustment have been shown to be associated with several psychiatric disorders. However, this flexible learning rate (FLR) account attributes all behavioral differences across volatility contexts solely to differences in learning rate. Here, we propose instead that different learning behaviors across volatility contexts arise from the mixed use of multiple decision strategies. Accordingly, we develop a hybrid mixture-of-strategy (MOS) model that incorporates the optimal strategy, which maximizes expected utility but is computationally expensive, and two additional heuristic strategies, which merely emphasize reward magnitude or repeated decisions but are computationally simpler. We tested our model on a dataset in which 54 healthy controls and 32 individuals with anxiety and depression performed a probabilistic reversal learning task with varying volatility conditions. Our MOS model outperforms several previous FLR models. Parameter analyses suggest that individuals with anxiety and depression prefer suboptimal heuristics over the optimal strategy. The relative strength of these two strategies also predicts individual variation in symptom severity. These findings underscore the importance of considering mixed strategy use in human learning and decision making and suggest atypical strategy preference as a potential mechanism for learning deficits in psychiatric disorders.

eLife assessment

This paper provides a **valuable** alternative explanation for the influence of environmental volatility on learning, attributing such effects to a mixture of strategies (MoS), rather than changes in the learning rate. The authors demonstrate that the MoS model provides a superior fit to previously published data, and suggest that atypical learning in individuals with anxiety and depression might reflect their use of a suboptimal strategy. While the approach should be of interest to researchers across decision sciences, the evidence is **incomplete**, limiting its potential impact.

Introduction

Intelligent behavior requires the ability to adapt to a constantly changing environment. For example, foraging animals must be able to track the changing abundance or scarcity of food resources in different locations and at different timescales. Motor control demands the ability to control limbs that constantly vary in their dynamics (due to fatigue, injury, growth, etc.). Human competitors in games or sports of all varieties must be able to learn and adapt to the changing strategies of their opponents.

To understand the mechanisms of these abilities, researchers have examined how (and how well) human agents can learn option values and track the dynamic changes in values in a volatile reversal learning task (Behrens et al., 2007). Unlike the traditional probabilistic reversal learning task where reward probabilities of two options only switch once (Cools et al., 2002), this paradigm includes two volatility conditions (see **Fig. 1B**): the reward probabilities of the two options keep constant in one condition (i.e., the *stable* condition) and switch periodically in the other (i.e., the *volatile* condition).

Previous studies often summarized human behaviors in this paradigm using the parameter of *learning rate*, a description of the efficiency with which current information is used to promote learning. The learning rate in essence serves as an abstract description of human learning behaviors, often exhibiting locality (Behrens et al., 2007; Boyd & Vandenberghe, 2004). Hence, the analyses of learning rates are usually contingent on the context, where researchers fit specific learning rates to each context. Using this method, previous studies have found that humans are able to flexibly adapt to the change in environmental volatility, which is exhibited by increasing and reducing the learning rate in response to volatile and stable conditions. Impaired flexibility in adjusting the learning rate according to environmental volatility has also been observed in individuals with several psychiatric diseases, including anxiety and depression (Behrens et al., 2007; Browning et al., 2015; Gagne et al., 2020). This hallmark can suggest atypical behaviors (Browning et al., 2015; Gagne et al., 2020), psychosis (Powers et al., 2017), and autism spectrum disorder (Lawson et al., 2017). Nevertheless, there are two limitations to this context-dependent method. First, as the number of contexts increases, the number of parameters can grow dramatically, increasing the risk of over-parameterization. Second, it can be challenging to interpret the learning rate in terms of its normative meaning. The quality of a learning rate never grows monotonically with its value but rather peaks at a moderate range. A higher learning rate is not always better. Apart from these, so far, there is no idea what the learning rate is associated with in the human brain.

The goal of the present work is to offer an alternative explanation of human reinforcement learning that is relatively context-independent. Instead of attributing the behavioral difference to the learning rate, we focus on a less-examined subcognitive process: *decision*. The decision process

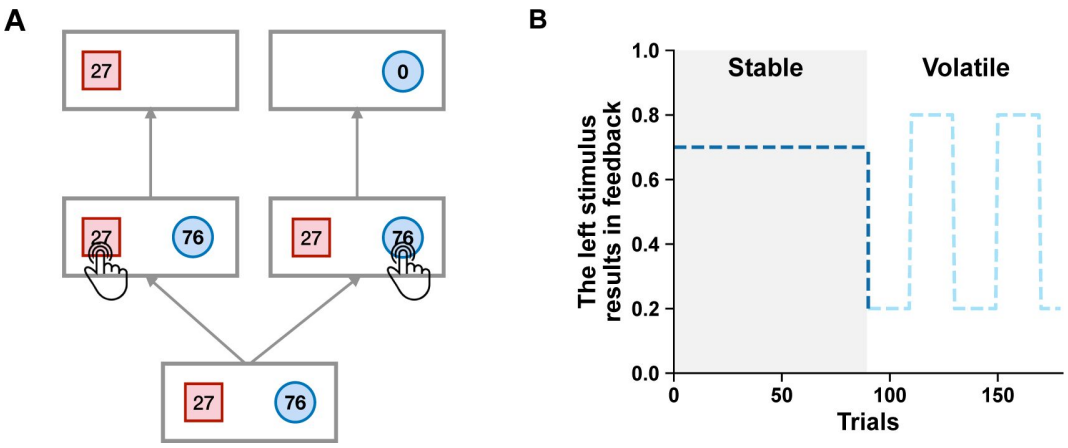


Figure 1

Schematic diagram of the experimental task in Gagne et al. (2020).

A. On each trial, the participants were shown two stimuli (with potential reward presented) and were instructed to choose one of them to receive feedback. Only one stimulus results in a reward.
B. Each task is evenly divided into two blocks: a stable and a volatile block. During the stable block, the environmental probability does not change, while in the volatile block, the probability flips every 20 trials.

describes how individuals strategically utilize their knowledge of the environment to generate responses. We realize this process by introducing a hybrid model, referred to as the *mixture-of-strategy* (MOS), which weights and sums over various strategies. The weighting parameters reflect subjects' decision preferences (Daw et al., 2011; Fan et al., 2023). As we will soon show, this model offers a parsimonious explanation for human behavioral responses across varying levels of environmental volatility, using a consistent set of weighting parameters across different contexts.

We base the MOS model on the principle of resource-rationality, which posits that humans' decision-making should tradeoff reward maximization and the consumption of cognitive resources (Gershman et al., 2015; Griffiths et al., 2015). Three strategies are included in the decision pool. First, we consider the optimal strategy, *Expected Utility* (EU), which guides decision-making based on the expected utility of each option (calculated as the probability multiplied by the reward magnitude) (Von Neumann & Morgenstern, 1947). This EU strategy yields the maximum amount of reward, but utility calculation *per se* consumes substantial cognitive resources. Alternatively, humans may choose simpler strategies. For example, the *magnitude-oriented* (MO) strategy, where only reward magnitude was considered during the decision process, and the *habitual* (HA) strategy, where people simply repeat choices frequently made in the past regardless of reward magnitude (Wood & Runger, 2016). These heuristic strategies certainly sacrifice the potential reward but come with the benefit of reducing cognitive costs in decision-making processes. We use the preference for these decision strategies to roughly estimate participants' reward-effort tradeoff in the volatile reversal task. Choosing to heavily rely on the EU strategy is more cognitively demanding than any combination of strategies. We expected that individuals with psychiatric diseases are less likely to use the EU strategy because they are known to have shrunk cognitive resources (Cohen et al., 2014; Harvey et al., 2005; Levens et al., 2009; Moran, 2016).

In this study, we apply and examine the MOS model on a dataset previously reported by Gagne et al. (2020). Our analysis reveals that, compared to healthy controls, patients with anxiety and depression exhibit a weaker tendency for the optimal EU strategy and a stronger preference for the simpler MO strategy, consistent with the reduced-resource hypothesis in psychiatric diseases. Furthermore, we demonstrate that this pattern of strategy preference readily accounts for several learning phenomena observed in prior research. Our work offers an alternative explanation for the effects of environmental volatility on human learning. Meanwhile, it underscores the importance of identifying behavioral markers to differentiate between explanations related to learning rate and decision strategy.

Methods and Materials

Datasets

We focused on the data from Experiment 1 reported in Gagne et al. (2020). The data is publicly available via (*https://osf.io/8mzuj/*). The original study included data from two experiments. The data from Experiment 2 was not used here because it was implemented on Amazon's Mechanical Turk with no information about the participants' clinical diagnoses. Here, we provide critical information about Experiment 1 (see Gagne, et al. (2020) for more technical details).

Participants

Eighty-six participants took part in this experiment. The pool includes 20 patients with a major depressive disorder (MDD), 12 patients with a generalized anxiety disorder (GAD), and 24 healthy control participants. The diagnosis was made through a phone screen, an in-person screening session, and the structured clinical interview following DSM-IV-TR (SCID). Thirty additional participants who reported no history of psychiatric or neurological conditions were recruited

without SCID. In this article, we regrouped the MDD and GAD individuals into a patient (PAT) group and the remaining 54 participants into a healthy control (HC) group. The detailed difference between MDD and GAD is not the focus of this paper. We will show later that the general factor behind MDD and GAD is the only factor that predicts learning behavior (see next section for details), a similar result reported in the original study (Gagne et al., 2020).

Clinical measures

The severity of anxiety and depression in all participants was measured by several standard clinical questionnaires, including the Spielberger State-Trait Anxiety Inventory (STAI form Y; Spielberger CD, 1983), the Beck Depression Inventory (BDI; Beck et al., 1961), the Mood and Anxiety Symptoms Questionnaire (MASQ; Clark & Watson, 1991; Watson & Clark, 1991), the Penn State Worry Questionnaire (Meyer et al., 1990), the Center for Epidemiologic Studies Depression Scale (CESD; Radloff, 2016), and the Eysenck Personality Questionnaire (EPQ; Eysenck & Eysenck, 1975). An exploratory bifactor analysis was then applied to item-level responses to disentangle the variance that is common to GAD and MDD or unique to each. The results of this analysis summarized participants' symptoms into three orthogonal factors: a general factor (g) explaining the common symptoms, a depression-specific factor (f1), and an anxiety-specific factor (f2). Similar to the original study, here we used the same three factors to indicate the participants' severity of their psychiatric symptoms.

Stimuli and behavioral task

This task is a volatile reversal learning task (see **Fig. 1A**). On each trial, participants were instructed to choose between two stimuli in order to receive feedback. There were two types of feedback. Participants received points or money in the reward condition and an electric shock in the punishment condition. The potential amount of reward or the intensity of electric shock (i.e., feedback magnitude) was presented together with the stimuli, but only one of the two stimuli would yield the feedback. The participant received feedback only after choosing the correct stimulus and received nothing else. The magnitude of the feedback, ranged between (1-99), is sampled uniformly for each shape from trial to trial. Each run consisted of 180 trials evenly divided into a stable and a volatile block (**Fig. 1B**). In the stable block, the dominant stimulus (i.e., the stimulus induces the feedback with a higher probability) provided a feedback with a fixed probability of 0.75, while the other one yielded a feedback with a probability of 0.25. In the volatile block, the dominant stimulus's feedback probability was 0.8, but the dominant stimulus switched between the two every 20 trials. Hence, this design required participants to actively learn and infer the changing stimulus-feedback contingency in the volatile block. The whole experiment included two runs each for the two feedback conditions. 79 participants completed both feedback conditions. 4 participants only completed the reward condition, and 3 participants completed the punishment conditions.

Computational Modeling

Each participant in the experiment must address two fundamental challenges: 1) decision-making, by adhering to a strategy that determines the action to maximize benefit; and 2) learning, by figuring out the untold feedback probability via feedback.

Before formalizing each challenge, we introduce our notation system. We denote each stimulus s as one of two possible states $s \in \{s_1, s_2\}$, s_1 refers to the left stimulus, and s_2 the right one. The labeled feedback magnitude (i.e., reward points or shock intensity) of the stimulus is $m(s)$, and the feedback probability is $\psi(s)$. Following the convention in reinforcement learning (Sutton & Barto, 2018), we presume that the decision is made from a policy π that maps the observed magnitudes m and currently maintained feedback probabilities ψ to a distribution over stimuli, $\pi(s|m, \psi)$. The construct of the policy varies between models (see below).

The mixture-of-strategy (MOS) model

The key signature of the hybrid MOS model is that its policy consists of a mixture of three strategies: expected utility (EU), magnitude-oriented (MO), and habitual (HA). The EU strategy postulates that human agents rationally calculate the value of each stimulus and use the softmax rule to select an action. In this case, the value of a stimulus should be its expected utility: $m(s)\psi(s)$.

The probability of choosing a stimulus s thus follows a softmax function.

$$\pi_{\mathrm{EU}}(s|\psi,m)=\frac{\exp\left(\beta\psi(s)m(s)\right)}{\sum_{s'}\exp\left(\beta\psi(s')m(s')\right)} \tag{1}$$

where β is the inverse temperature that is used to round the policy to a Bernoulli distribution. For simplicity, we rewrite Eq. 1 in the following form:

$$\pi_{\mathrm{EU}}(s|\psi,m)=\mathrm{softmax}\left(\beta\psi(s)m(s)\right) \tag{2}$$

Different from the EU strategy, the MO strategy postulates that observers only focus on feedback magnitude $m(s)$, disregarding feedback probability $\psi(s)$. This is certainly an irrational strategy but more economical in terms of cognitive efforts. Feedback magnitudes are explicitly shown with the stimuli in each trial and readily available for any related cognitive computation. But feedback probability, as a latent variable, requires trial-by-trial learning and inference, which is more cognitively demanding. The MO strategy is defined as,

$$\pi_{\mathrm{MO}}(s|m)=\ \mathrm{softmax}\left(\beta m(s)\right) \tag{3}$$

Like the EU strategy, the MO strategy is converted to a Bernoulli distribution after passing through a softmax function. This processing is necessary because a hybrid model is uninterpretable when its components follow heterogeneous distributions. The softmax function enhances the model's interpretability.

Unlike EU and MO, the HA strategy depends on neither feedback magnitude $m(s)$ nor feedback probability $\psi(s)$. The HA strategy reflects the tendency to repeat previous frequent choices. This reinforcement reflects the habit of choosing a stimulus, a phenomenon sometimes called hot-hand bias (Gilovich et al., 1985) or perseveration (Gershman, 2020; Wood & Runger, 2016) in literature. For example, if an agent chooses the left stimulus more often in past trials, she forms a preference for the left stimulus in future trials. We constructed it as a Bernoulli distribution (henceforth called habitual distribution) over the two stimuli $\pi_{\mathrm{HA}}(s)$. The trial-by-trial update rule of $\pi_{\mathrm{HA}}(s)$ will be detailed in Eqs. 5–6 below.

We implemented the hybrid policy of a linear mixture of the three strategies following the methods used in Daw et al. (2011),

$$\pi(s|\psi,m,\pi_{\mathrm{HA}})=w_{\mathrm{EU}}\pi_{\mathrm{EU}}(s|\psi,m)+w_{\mathrm{MO}}\pi_{\mathrm{MO}}(s|m)+w_{\mathrm{HA}}\pi_{\mathrm{HA}} \tag{4}$$

where w_{EU}, w_{MO}, and w_{HA} are the weighting parameters of each strategy. The three weighting parameters should be summed to 1, i.e., $w_{\mathrm{EU}}+w_{\mathrm{MO}}+w_{\mathrm{HA}}=1$. We can thus describe the policy an observer adopted just by examining the weighting parameters.

Next, we modeled the second challenge — the probabilistic learning process. Two distributions — the feedback probability and the habitual probability — are learned and updated in a trial-by-trial fashion. We updated the feedback probability according to the outcome of the left stimulus s_1:

$$\begin{aligned}\psi(s_1) &= \psi(s_1) + \alpha_\psi\big(O(s_1) - \psi(s_1)\big)\\ \psi(s_2) &= 1 - \psi(s_1)\end{aligned} \tag{5}$$

where α_ψ is the learning rate. $O(\cdot)$ is an indicator function that returns 1 at the true feedback stimulus or 0 otherwise. Intuitively, the stimulus that induced a reward would be reinforced and its feedback probability was enhanced. This update equation is the standard format of the well-known Rescorla-Wagner model (Rescorla, 1972). To keep consistent with Gagne et al., (2020), we also explored the valence-specific learning rate in some models,

$$\alpha_\psi = \begin{cases}\alpha_{\psi+}, & \text{for } (O(s_1) - \psi(s_1)) > 0\\ \alpha_{\psi-}, & \text{for } (O(s_1) - \psi(s_1)) < 0\end{cases}$$

The habitual distribution is updated in a similar manner.

$$\begin{aligned}\pi_{\mathrm{HA}}(s_1) &= \pi_{\mathrm{HA}}(s_1) + \alpha_\pi\big(A(s_1) - \pi_{\mathrm{HA}}(s_1)\big)\\ \pi_{\mathrm{HA}}(s_2) &= 1 - \pi_{\mathrm{HA}}(s_1)\end{aligned} \tag{6}$$

where α_π is the learning rate. $A(\cdot)$ is also an indicator function that returns 1 for the stimulus chosen at the current trial. Intuitively, the stimulus chosen in each trial regardless of its feedback will be reinforced via Eq. 6.

We developed two variants for each model, a context-free and a context-dependent variant. The context-free MOS6 has six parameters $\xi = \{\beta, \alpha_{\mathrm{HA}}, \alpha_\psi, w_{\mathrm{EU}}, w_{\mathrm{MO}}, w_{\mathrm{HA}}\}$. This variant does not include the value-specific learning rate design. The context dependent variant MOS22 has 22 parameters. Among them β and α_{HA} are context-free parameters that holds the same for all contexts. Parameters $\{\alpha_{\psi+}, \alpha_{\psi-}, w_{\mathrm{EU}}, w_{\mathrm{MO}}, w_{\mathrm{HA}}\}$ are context-dependent parameters that are fitted specifically to each context. We will further discuss the model fitting details in the later section.

The flexible learning rate (FLR) model

The FLR model refers to Model 11 (i.e., the best-fitting model) in Gagne et al. (2020). Here, we describe the FLR model using the same notation system with the published paper, slightly different from the notations in the MOS model. The FLR model models the probability of selecting the left stimulus s_1 as,

$$\pi(s_1|v, \pi_{\mathrm{HA}}) = \frac{1}{1 + \exp(-\beta v - \beta_{\mathrm{HA}}[\pi_{\mathrm{HA}}(s_1) - \pi_{\mathrm{HA}}(s_2)])} \tag{7}$$

where β and β_{HA} are the inverse temperature parameters of the value of the left stimulus and the HA strategy, respectively. The value of the left stimulus v represents the advantage of s_1 over s_2,

$$v = \lambda[\psi(s_1) - \psi(s_2)] + (1 - \lambda)\mathrm{sign}\big(m(s_1) - m(s_2)\big)|m(s_1) - m(s_2)|^r \tag{8}$$

where λ is the weighting parameter balancing the two terms. The first term $\psi(s_1) - \psi(s_2)$ indicates the feedback probability difference between the two options. The second term, $\mathrm{sign}(m(s_1) - m(s_2))\,|m(s_1) - m(s_2)|^r$, indicates the feedback magnitude differences scaled by a non-linear factor r. Intuitively, the value v of s_1 can be understood as the weighted sum of the two terms. We write this nonlinear scaling in a slightly different form with Eq. 1b in Gagne et al. (2020) to better replicate their coding implementation.

During the learning stage, the FLR model learns the feedback probability using the same equations in the MOS model (Eqs. 5–6). The context-free variant FLR6 has 6 parameters $\xi = \{\alpha_{HA}, \beta_{HA}, r, \alpha_{\psi}, \beta, \lambda\}$. The context-dependent variant FLR19 considers $\{\alpha_{HA}, \beta_{HA}, r\}$ as context-free parameters; $\{\alpha_{\psi+}, \alpha_{\psi-}, \beta, \lambda\}$ as context-dependent parameters.

The risk-sensitive (RS) model

We adopted the RS model from Behrens, et al. (2007). The RS model assumes that participants apply the EU policy but with a subjectively distorted feedback probability $\tilde{\psi}(s_1)$,

$$\pi\left(s_1|\tilde{\psi}, m\right) = \frac{1}{1+\exp(-\beta[\tilde{\psi}(s_1)m(s_1) - \tilde{\psi}(s_2)m(s_2)])} \tag{9}$$

where β is the inverse temperature. The distorted probability is calculated by,

$$\begin{aligned} \tilde{\psi}(s_1) &= \max[\min[(\gamma(\psi(s_1) - 0{,}5) + 0.5, 1], 0] \\ \tilde{\psi}(s_2) &= 1 - \tilde{\psi}(s_1) \end{aligned} \tag{10}$$

where the γ indicates participants' risk sensitivity. When $\gamma = 1$, a participant has an optimal risk balance. $\gamma < 1$ and $\gamma > 1$ indicate risk-seeking and risk-aversive tendencies, respectively.

The RS model learns the feedback probability the same as the MOS and FLR models (i.e., Eq. 5). The model did not include the HA strategy. The context-free variant RS3 has 3 parameters $\xi = \{\beta, \alpha_{\psi}, \gamma\}$. The context-dependent variant RS12 considers $\{\beta, \alpha_{\psi+}, \alpha_{\psi-}, \gamma\}$ as context-dependent parameters.

The Pearce-Hall (PH) model

That people use different learning rates in different contexts implies that people adaptively adjust the learning rate during the learning process. To constitute this hypothesis, we adopt the PH model, an adaptive learning rate model, from Pearce and Hall (1980). The PH model posits that an adaptive learning rate in Eq. 5,

$$\begin{aligned} \psi(s_1) &= \psi(s_1) + k\alpha_{\psi}\left(O(s_1) - \psi(s_1)\right) \\ \psi(s_2) &= 1 - \psi(s_1) \end{aligned} \tag{11}$$

where k is a scale factor of the learning rate. Each trial the learning rate is updated in accordance with the absolute prediction error,

$$\alpha_{\psi} = \alpha_{\psi} + \eta(|O(s_1) - \psi(s_1)| - \alpha_{\psi}) \tag{12}$$

where η is the step size for the learning rate. We have no knowledge of participants' learning rate values before the experiment, so we need to also fit the initial learning rate value, α_{ψ}^0. The PH model generate a choice through a sigmoid function,

$$\pi(s_1|\psi, m) = \frac{1}{1+\exp(-\beta[\psi(s_1)m(s_1) - \psi(s_2)m(s_2)])} \tag{13}$$

The context-free variant PH4 has 4 parameters $\xi = \{\alpha_{\psi}^0, k, \eta, \beta\}$. The context-dependent variant PH17 considers $\{\alpha_{\psi}^0\}$ as context-free parameters; $\{k_+, k_-, \eta, \gamma\}$ as context-dependent parameters.

Model fitting

To characterize participants' behavioral patterns in different experimental context c, we fit the context-dependent parameters to each context following a 2-by-2 factorial structure (Table 1). For example, in the MOS model, we were only interested in the learning rate parameters α_ψ and three strategies weights w_{EU}, w_{MO}, w_{HA} and fit them separately to each context. The remaining two parameters $\{\beta, \alpha_{\mathrm{HA}}\}$ were held constant across all four experimental contexts for each participant. Thus, there were 22 free parameters (2 context-free parameters + 5 context-dependent parameters × 4 conditions) of the MOS model in each participant. In contrast, the context-free variant (MOS6), we fit the same set of parameters to all contexts.

Parameters were estimated for each participant via the maximum a posteriori (MAP) method. The objective function to maximize is:

$$\max_{\xi(c)} \sum_{i=1}^{N} \log L\left(s_i \middle| m_i, O_i, \mathrm{M}, \xi(c)\right) + \log p\left(\xi(c)\right) \tag{11}$$

where $\xi(c)$ means the model parameters under condition c. M is the model and N refers to the number of trials of the participant's behavioral data in condition c. m_i, O_i, and s_i are the presented magnitude, true feedback probability, and participants' responses recorded in each trial.

Parameter estimation was performed using the *Broyden-Fletcher-Goldfarb-Shanno (BFGS)* algorithm in the *scipy.optimize* module in Python. This algorithm provides an approximation of the inverse Hessian matrix for the parameter, a critical component that can be employed in Bayesian model selection (Rigoux et al., 2014). For each participant, we ran the optimization with 40 randomly chosen initial parameters to avoid local minima.

In order to use the BFGS algorithm, we reparametrized the model, thereby transforming the original fitting problem into an unconstrained optimization problem (Supplemental material Note 1). Importantly, to fit the weighting parameters (w_{EU}, w_{MO}, w_{HA}) and ensure them summed to 1, we parameterized the weighting parameters as outputs of a softmax function,

$$w_i = \mathrm{softmax}(\lambda_i) \qquad \forall i \in \{\mathrm{EU}, \mathrm{MO}, \mathrm{HA}\} \tag{14}$$

and fit the logits λ_i of the weights. All logits were assumed to be normally distributed with a prior N (0, 10). Due to its normality, we also used the logits as participants' strategy preferences in statistical analyses in the result section. To provide better intuition, we will use the terms "weighting parameters" and "logits" interchangeably in the following section. Specifically, we may refer to the logits λ_{EU}, λ_{MO}, and λ_{HA} as the weighting parameters.

Simulate to explain the previous learning rate effects

In the result section, *Explain the previous learning rate effects using the strategy preferences*, we illustrate how we explicate some classical learning rate phenomena on the weighting parameters of the MOS model, referred to as the strategy preferences. Here are some technical details.

We simulate to show that the strategy preferences alone can explain the slower learning curve in the patient group. Each simulation task included 90 stable trials, followed by 90 volatile trials. The parameters used for simulations are β = 8.536, α_{HA} = 0.403, α_ψ = 0.460, λ_{EU} = 0.712, λ_{MO} = −0.988, λ_{HA} = 0.276. We outputted the predicted probability of the left action a_1 for each strategy along with the learning trials to generate **Fig. 4B**. We simulated the policy for the healthy control group using the averaged parameters except for replacing the three weighting parameters with $\lambda_{\mathrm{EU}}^{\mathrm{HC}} = 0.873$, $\lambda_{\mathrm{MO}}^{\mathrm{HC}} = -1.575$, $\lambda_{\mathrm{HA}}^{\mathrm{HC}} = 0.701$ (**Fig. 4A**, HC curve). The same method was applied HC HC to generate the PAT curve using parameters $\lambda_{\mathrm{EU}}^{\mathrm{PAT}} = 0.109$, $\lambda_{\mathrm{MO}}^{\mathrm{PAT}} = -0.072$, $\lambda_{\mathrm{HA}}^{\mathrm{PAT}} = -0.037$.

	Reward	Aversive
Stable	Reward-Stable	Aversive-Stable
Volatile	Reward-Volatile	Aversive-Stable

Table 1

Four experimental contexts

Demonstrating the remaining two effects is equivalent to establishing the following assertion: when compared to the MO strategy, a preference for the EU strategy results in a qualitatively greater extent of increase from stable to volatile conditions, whereas a preference for the MO strategy corresponds to a lesser increase extent. We generated 10 blocks of synthesized data on the MOS model with reward feedback using $\beta = 8.536$, $\alpha_{HA} = 0.403$, $\alpha_{\psi} = 0.460$, 5 blocks each for the EU and MO strategies. For the EU strategy, we set weighting parameters $\lambda_{EU} = 10$, $\lambda_{MO} = 0$, $\lambda_{HA} = 0$, which yields $w_{EU} \approx 1$; Similarly, we set the weighting parameters to $\lambda_{EU} = 0$, $\lambda_{MO} = 10$, $\lambda_{HA} = 0$ to synthesize the learning curve for the MO strategy, resulting in $w_{MO} \approx 1$. We then fit the FLR and the RS models to these synthesized data, controlling all parameters except for the learning rate.

All parameter values introduced here are reparametrized rather than the raw values.

Parameter recovery and model recovery analyses

We conducted a parameter recovery analysis to validate the fitting of the MOS models. We generated 80 synthetic datasets varying the four parameters of interest $\{\alpha_{\psi}, \lambda_{EU}, \lambda_{MO}, \lambda_{HA}\}$. The remaining parameters were fixed to the averaged fitted weighting parameters $\beta = 8.804$, $\alpha_{HA} = 0.366$. Each dataset in both cases contained ten blocks. For each dataset, we fit our MOS6 model to the data and compared the fitted parameters and the ground-truth parameters. Parameter recovery analysis aims to exclude the exchangeability between the learning rate and weighting parameters.

To further differentiate models, we also performed a model recovery analysis. We generated 40 synthetic ten-block datasets from the MOS6 model, using the parameters fitted to each participant. We fit all six models to each dataset and examined whether the MOS6 model, as the generative model, was still the best-fitting model on the synthetic datasets.

Results

To illustrate that the mixture of strategies provides parsimonious alternative explanations, we first demonstrate that the context-free MOS6 model can quantitatively capture human learning behaviors and predict individual psychiatric syndromes. Furthermore, we simulate to show that the MOS6 model, without its parameters held constant, can explain certain human behavioral phenomena, as previously evidenced by context-dependent learning rates.

The mixture-of-policy model quantitatively captures learning behaviors

We fit a total of eight models to the behavioral data reported in Gagne et al. (2020). Model fitting and comparison results are summarized in Table 2. To quantify goodness-of-fit, we calculated negative total likelihood (NLL), Akaike Information Criterion (AIC; Akaike, 1974), and Bayesian Information Criterion (BIC; Schwarz, 1978) for each individual participant and performed Bayesian model selection at the group level (Rigoux et al., 2014).

Modeling fitting reveals that the MOS framework accurately accounts for human behaviors. MOS6 and MOS22 were the best-fitting models in terms of BIC and AIC, respectively. The different results based on AIC and BIC may be due to the different degrees of penalty on model complexity (i.e., number of free parameters). The group-level Bayesian model comparisons suggested MOS6 as the best-fitting model. These model comparisons underscore that the MOS framework outperforms existing models, FLR and RS, in the previous literature (Behrens et al., 2007; Gagne et al., 2020). Importantly, an analysis of the parameters in the MOS22 model revealed no significant

differences across different experimental contexts (discussed later). This suggests that MOS22 and MOS6 are not qualitatively distinct. Therefore, we conclude that the MOS6 model can effectively account for human behaviors in a relatively context-free manner.

The difference in learning rates between stable and volatile conditions highlights the human capacity to flexibly adapt their learning rate in response to environmental volatility. To explore this adaptability, we applied a model with a built-in adaptive learning rate known as the Pearce-Hall model (PH4, PH17). However, our findings indicate that this model does not provide a better explanation than the MOS6 model. This suggests that there may be behavioral variations that cannot be fully accounted for by the parameters of learning rate.

MDD and GAD patients favor simpler and more irrational strategies

The MOS model assumes that each participant's response is a result of a weighted combination of three strategies, EU, MO, and HA. We therefore can summarize participants' decision preferences using the weighting parameters w. For example, a larger weight w_{EU} for the EU strategy indicates the participant's tendency of using the rational strategy in value computation and action selection. For the significant testing, we used the logit (λ) of the weight parameters (w) as indicators of decision preference for significant testing. This is because the weighting parameters are not normally distributed, but their logits are approximately subject to the normal assumption. We performed a *Welch's t-test* (Delacre et al., 2017) on the MOS6 model to ensure a reliable analysis of the unequal population for the health control and patient groups. We found that the patient group exhibited a weaker tendency for the rational EU strategy ($t(57.980) = 2.195$, $p = 0.032$, Cohen's d = 0.508) and the HA strategy ($t(59.032) = 2.389$, $p = 0.020$, Cohen's d = 0.550), but stronger tendency for the MO strategy($t(63.746) = -3.479$, $p = 0.001$, Cohen's d = 0.783) (**Fig. 3A**). However, there was no group difference in (log) learning rate ($t(72.041) = 0.678$, $p = 0.500$, Cohen's d = 0.147).

For completeness, we also examined whether the decision preferences and the learning rates varied as a function of volatility levels (stable/volatile) and feedback types (reward/aversive) using the MOS22. We conducted three 2 × 2 × 2 ANOVAs on the logit of the three weighting parameters of MOS22 and found no significant relationship between different volatility levels or feedback types (all ps > 0.149) except that participants are more likely to use EU strategy under the reward condition ($F(1, 300) = 13.426$, $p = 0.021$, $\eta^2 = 0.016$; see more details in Supplemental Note 2). In addition, the between-group (HC/PAT) analyses of the decision preferences in MOS22 are mostly consistent with the MOS6 results shown above (**Supplemental Fig. S2**). For the (log) learning rate parameters, we also examined the outcome valence (higher/lower reward than expectation) apart from the volatility levels or feedback types, finding no general significance (all ps > 0.046; **Supplemental material Note 2**). All these results suggest that one set of parameters is sufficient for the MOS model to describe the human behavioral dataset.

Decision preferences predict the general severity of anxiety and depression

We investigated the relationship between decision preferences and psychiatric symptom severity (**Fig. 3B**). To measure symptom severity, we used the bifactor analysis approach described by Gagne et al., (2020) which decomposed measurements of symptom severity into the factors specific to anxiety and depression, and the general factor (g score) indicating the common symptoms shared by them. Our findings indicate that patients with severe symptoms exhibit a weaker tendency of using the optimal EU strategy (Pearson's $r = -0.221$, $p = 0.040$) but a stronger tendency of using the MO strategy (Pearson's $r = 0.360$, $p = 0.001$). Additionally, there was a significant correlation between symptom severity and the preference for the HA strategy (Pearson's $r = -0.285$, $p = 0.007$). In other words, the participants with severer anxiety tend to use a

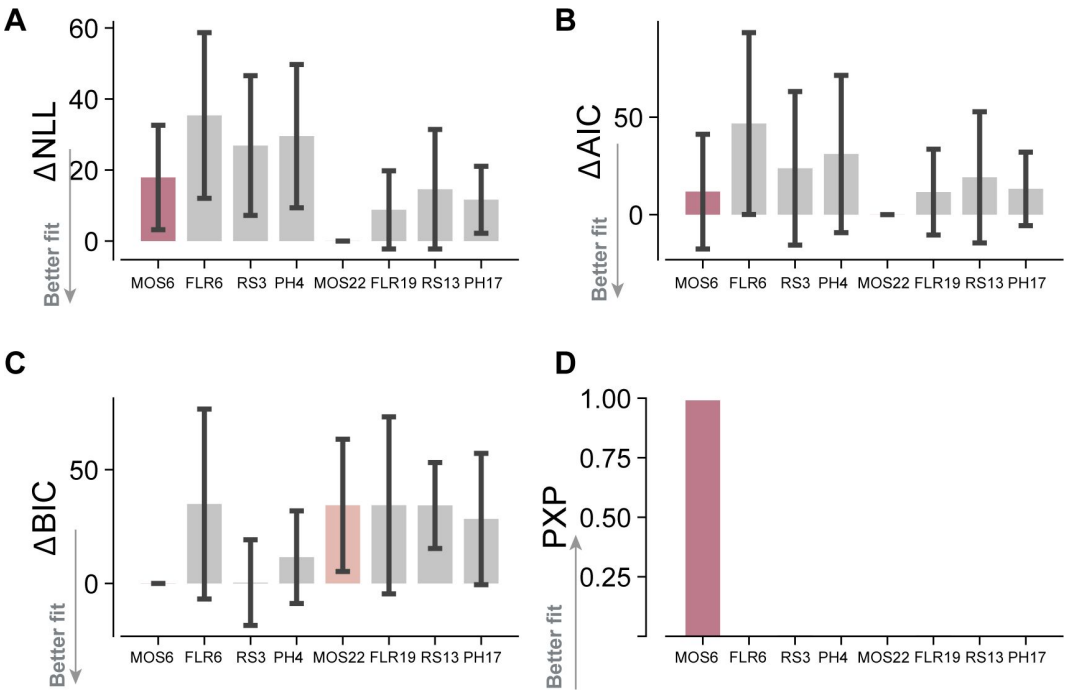


Figure 2

Model comparison results. The reds indicate the target models.

A-C. Averaged relative increase in negative log-likelihood (NLL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) subtracted by the lowest values. Lower scores indicate better models. Error bars indicate the standard deviation for the estimated mean across 86 participants.
D. Protected exceedance probability (PXP) for the group-level Bayesian model selection.

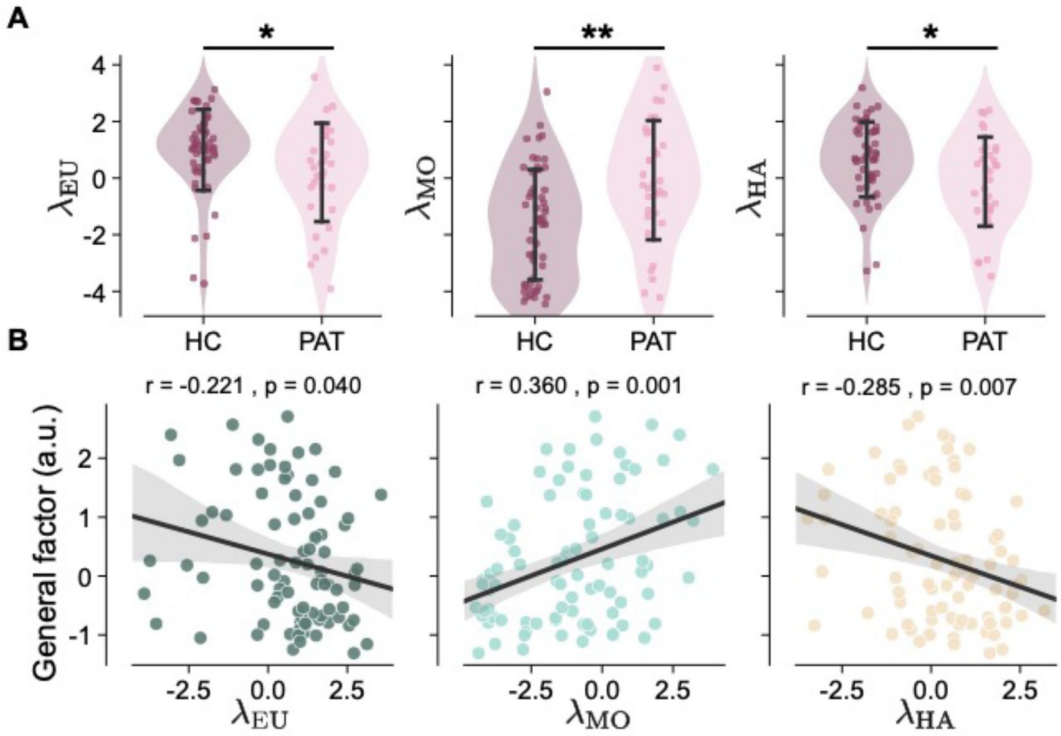


Figure 3

Parameter analyses of the MOS6 model.

A. The weighting parameters for the healthy control participants (HC, purple) and the patients (PAT, pink) diagnosed with MDD and GAD. The y-axis means averaged preference over different volatility levels (volatile and stable) and feedback types (reward and aversive). Error bars reflect the standard deviation across 86 participants. Significance symbol conventions are: *: $p < 0.05$; $p < 0.01$; $p < 0.001$; n.s.: non-significant.

B. Decision preferences predict participants' general factor score (*g* score) in the bifactor analysis reported in Gagne, et al. (2020). The y-axis indicates the averaged preference over different volatility levels (volatile and stable) and feedback types (reward and aversive). This average operation is permitted here because the logit of the weight is normally distributed. The shaded areas reflect 95% confidence intervals of the regression prediction.

less accurate but simpler strategy for probabilistic learning. Again, we suspect that this is because anxiety and depression reduce cognitive resources in patients, and they have to choose less resource-consuming strategies. We will return to this point in the discussion.

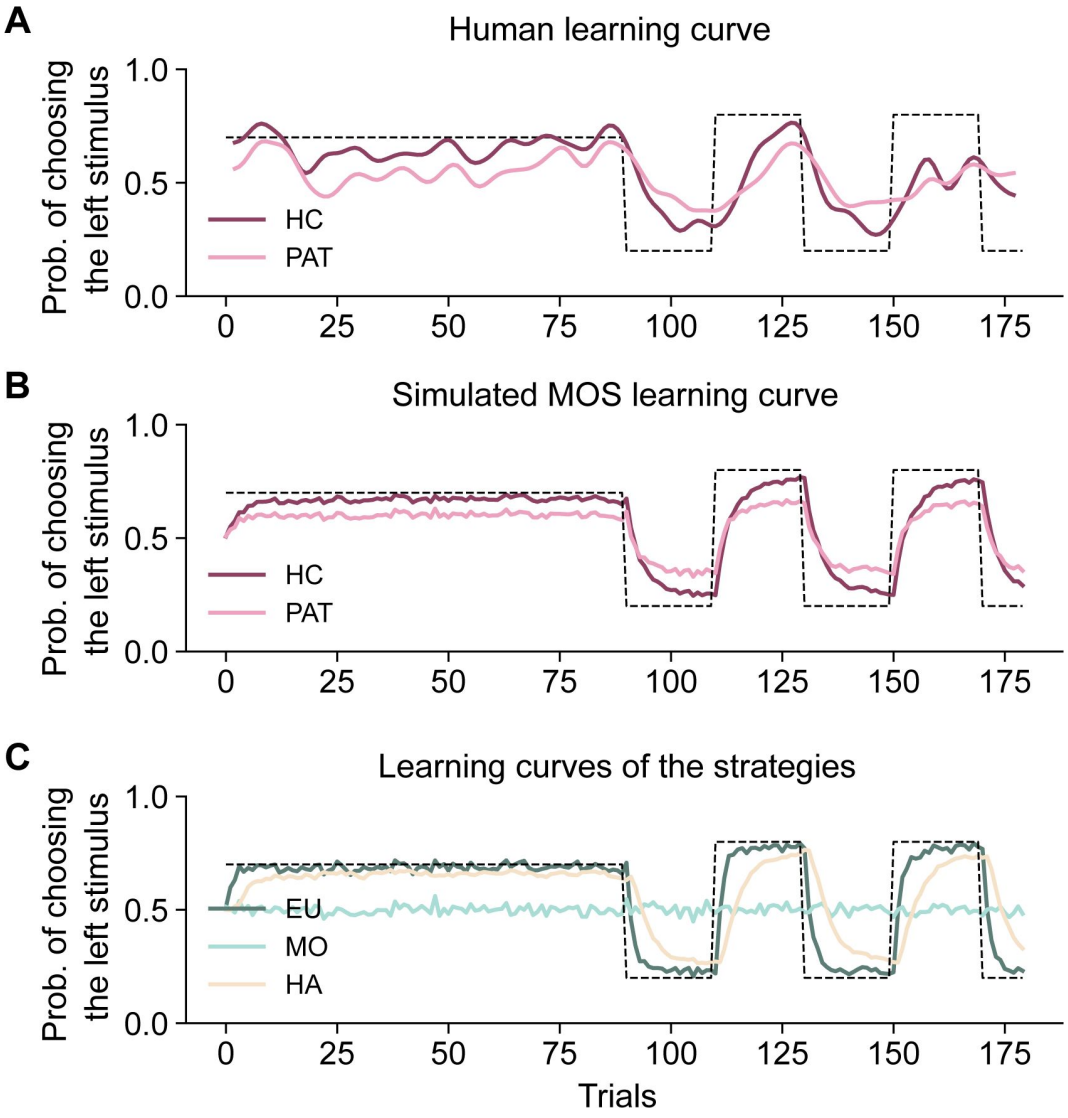
Explain learning rate effects using the strategy preferences

Three ubiquitous observations have been documented in probabilistic learning tasks. First, individuals with anxiety and depression often exhibit a slower learning curve over the course of learning, as evidenced by a smaller fitted learning rate (Chen et al., 2015; Pike & Robinson, 2022) (Chen et al., 2015). Second, to adapt to the high environmental volatility, subjects increase their learning rate to generate a faster learning curve (Behrens et al., 2007). Third, the extent of the learning rate increment from the stable to the volatile condition is smaller in the patient group, a hallmark of their learning deficits (Browning et al., 2015; Gagne et al., 2020). Here, we demonstrate that the MOS model can qualitatively reproduce all three effects by only attributing them to strategy preferences without resorting to the learning rate parameter.

We first averaged over data from 43 human participants (26 health control and 17 patients) in an experimental context and showed that the patients exhibited slower convergence to the true feedback probability than the healthy control group (**Fig. 4A**). Next, we used the MOS model to simulate the learning behaviors of the two groups using the averaged weighting parameters $\{w_{EU}, w_{MO}, w_{HA}\}$ of each group. Meanwhile, we controlled the learning rate effect by fixing the parameters $\{\beta, \alpha_{HA}, \alpha_{\psi}\}$ to their averaged values across all participants for both HA and PAT groups, volatility levels (stable/volatile), and feedback types (reward/punishment).

Without introducing the learning rate difference between healthy controls and patients, we observed the same slower learning curve in the patient group in the simulations (**Fig. 4B**). To gain insights, we visualized their respective learning curves throughout the learning task (**Fig. 4C**). The EU strategy, which is theoretically optimal, quickly approximates the true feedback probability and exhibits a faster learning curve. The HA strategy can also adapt to the volatile environment at a slower adaptation speed and with longer delays, resulting in a slower learning curve. This is intuitively reasonable because shaping a habit usually takes a longer time. The MO strategy is not adaptive to environmental volatility at all and exhibits a flat learning curve throughout the entire course of learning. As previously mentioned, the patients tend to be more magnitude-oriented (MO) possibly because of their inability to afford the effort-consuming EU strategy. Their preference for the slowest decision strategy induces a flattened learning curve in the probabilistic learning task. Therefore, we conclude that strategy preference can explain the slower learning curves in the patient group.

Next, we investigate whether the strategy preferences can account for the remaining two effects. Based on our earlier conclusion that health participants prefer the EU strategy while patients favor the MO strategy, we equate this problem as showing that a preference for the EU strategy results in a greater extent of increase in the learning rate from stable to volatile condition, whereas a preference for the MO strategy corresponds to a lesser extent of increase. To this end, we fit the FLR and RS models to the simulated data generated by each strategy in the MOS model. We controlled all parameters except for the learning rate parameters across the two strategies (see Methods **Simulate to explain the previous learning rate effects** for details). We observed that, for the EU strategy, there was an elevation in the fitted learning rate from the stable to the volatile condition (**Fig. 5 Learning rate**), which mirrors the findings of faster learning curves in the volatile condition. The MO strategy displayed almost no increase. Furthermore, we also found that the increase in the learning rate was smaller for the MO strategy (**Fig. 5 Learning rate: volatile - stable**), indicating that patients would display a smaller increase in the learning rate. These results suggest that strategy preferences alone can provide a natural explanation for patients' maladaptive learning behaviors in response to environmental volatility.



C. The simulated learning curve for each strategy.

Figure 4

Simulated learning behavior of the two groups and the three strategies.
The black dashed lines indicate the ground truth feedback probability.
The simulated curves in B-C were generated by averaging 500 simulations.

A. The human learning curves of the two groups were produced using the data in the aversive context (see the reward context in **Fig. S3**) and smoothed by a Gaussian kernel with a window size of 5 trials and a s.t.d. of 2 trials.
B. Simulated the learning curves of the two groups using their fitted parameters in the MOS model.

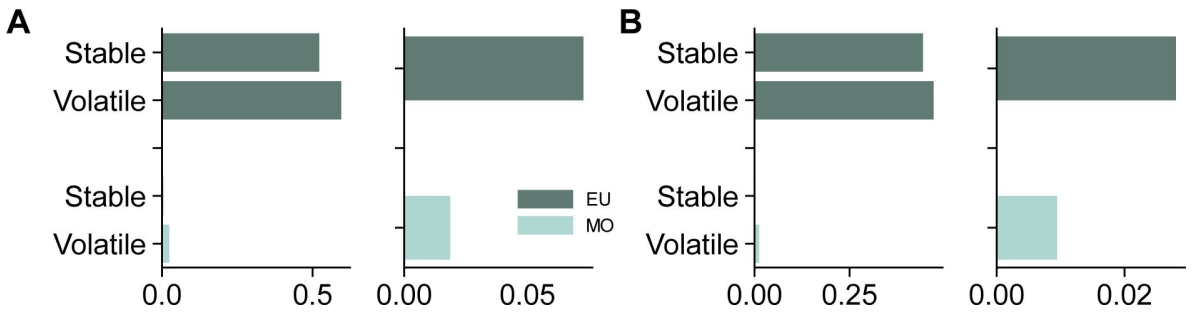


Figure 5

Fitted learning rate and the learning rate differences between the stable and volatile conditions. The simulated data are generated by the MOS model and fitted by the FLR (*A*) and the RS (*B*) model.

In summary, the MOS model can effectively explain the three well-established learning curve effects in previous literature. It is important to note that, in contrast to the FLR or RS models, the apparent differences in learning curves in the MOS model originate from the weighting differences in strategy rather than learning rate *per se*. This means that the MOS model provides a key theoretical interpretation that differs from that in the majority of literature.

Model and parameter recovery analyses support model and parameter identifiability in MOS

It is intriguing that the MOS model can reproduce the classic learning curve effects only by adjusting strategy preferences without altering the learning rates. However, there are two potential confounding factors to consider. First, it is possible that adjusting the learning rate, rather than strategy preferences, could produce the same behavioral outcomes that are indistinguishable by the model fitting. If this holds, the MOS model might be problematic, as all learning rate differences may be automatically attributed to strategy preferences because of some unknown idiosyncratic model fitting mechanisms. Second, the fact that the MOS framework outperforms the other two frameworks may be partly due to an unknown bias in the model design. It is possible that the MOS model always wins, irrespective of how the data is generated.

To circumvent these issues, we performed parameter and model recovery analyses to investigate the identifiability of true parameters and models. The parameter recovery results demonstrate that the true parameters that generate synthetic datasets can be correctly estimated and identified (all Pearson's rs > 0.720), demonstrating that the effects of learning rate and weighting parameters are not interchangeable in the MOS6 model.

For model recovery, we fit all six models to the synthetic data generated by MOS6 and found MOS6, as the generative model, was still the best-fitting model based on the lowest averaged AIC and BIC (**Fig.** 7). Both parameter and model recovery analyses suggest that our modeling approach is reliable and the MOS6 model is distinguishable. We ensured that differences in decision preferences between patients and HC were not the result of idiosyncratic model design or fitting procedures. Remark that we excluded the NLL and PXP from the evaluation. The NLL always favors models with more parameters, while the use of PXP, which is designed for group-level comparisons, is not an appropriate metric here since we knew in advance that all data were generated from one identical model in this model recovery.

Discussion

In this article, we propose a mixture-of-strategy model assuming that human agents' decision policy consists of three distinct components: the EU, the MO, and the HA strategies. The EU strategy is optimal in terms of maximizing reward. The MO and HA are simpler heuristic strategies that are cognitively demanding. We applied the MOS model to a public dataset and found that it outperformed existing models in capturing human behaviors. We summarized human behaviors using the estimated parameters of the target model and reported three primary conclusions. First, individuals with MDD and GAD tended to favor more irrational policies (i.e., a stronger preference for the MO strategy). Second, individual decision preferences predict the general severity of anxiety and depression. Third, decision preferences explain several learning rate phenomena that have been studied before. All conclusions suggest that a mixture of strategies provides an effective and parsimonious explanation for human learning behaviors in volatile reversal tasks.

The attempt of decision analysis in the previous studies

We are not the first to examine the human decision process. Numerous previous studies have also explored this cognitive process, although they did produce particular findings.

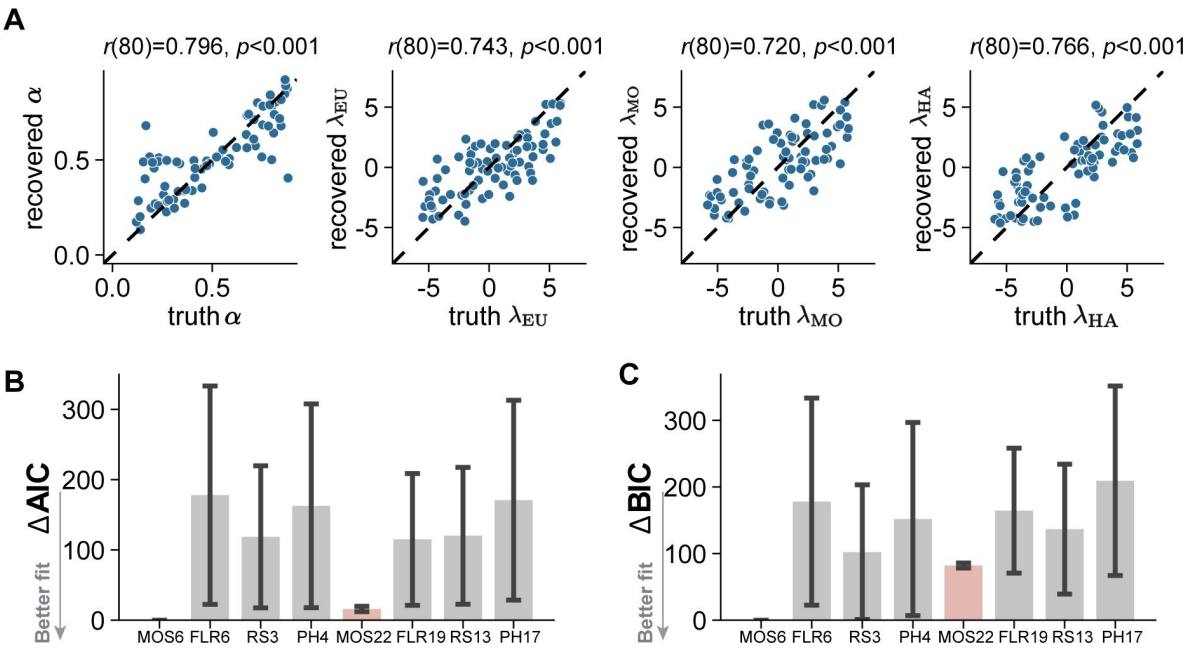


Figure 6

Parameter and model recovery analyses

A. Parameter recovery for the MOS6 model. Each recovered parameter is averaged over ten samples.
B-C. Model recovery analyses. The MOS6 model is still the best-fitting model on synthetic data generated by MOS6 *per se*, according to the AIC (*B*) and BIC (*C*) comparisons. Error bars indicate the standard deviation of the mean value across 79 synthetic data points.

The well-established finding that humans apply flexible learning rates in different experimental blocks is a successful case study of the ideal observer analysis. Behrens et al. (2007) constructed a hierarchical ideal Bayesian observer that dynamically models how higher-order environmental volatility influences the speed of updating the lower-order feedback probability. Because of the hierarchical interaction, the model predicts a faster updating speed for the feedback probability in a volatile environment. The ideal Bayesian model proposes an optimal manner of processing new information, like how an agent should behave. Human behavioral data was better accounted for by the RS model, which updates feedback probability in the classical Rescorla-Wagner format. Interestingly, the key prediction of the ideal Bayesian model has been preserved in the RS implementation: human subjects had a significantly higher learning rate in the volatile than the stable environment. The success of the RS model seems to suggest that humans can flexibly adjust learning rates according to environmental volatility. The view is better established as more studies replicate this learning rate effect (e.g., Browning et al., 2015; Gagne et al., 2020).

Despite this, some attention has nevertheless been paid to understanding the decision process. Browning et al. (2015) studied the decision process of the RS9 model. They examined and compared the risk-sensitive parameter γ and inverse temperature β but found no significant difference between different degrees of trait anxiety and volatility. Gagne et al. (2020) constructed 13 models in a stepwise manner to find the best-fitting description of human decision-making in the volatile reversal learning task. However, the study did not attempt to connect the decision process to anxiety and depression traits, possibly due to the best-fit model, the FLR18 model implemented here, being too complex to analyze.

The problems in both attempts are straightforward. The RS9 model might have an inaccurate description of the human decision process, and the FLR18 model is not understandable. The MOS model developed here relieves both issues, providing a competitive fit and being constructed in an easy-to-understand form. Additionally, the MOS model also provides a parsimonious description of the behavioral data, using a set of parameters to capture the data in four experimental contexts. However, the model yields a contradictory explanation. It shows no significant difference in terms of learning rate. In other words, the apparent differences in learning curves may arise from decision processes (i.e., decision preferences) rather than learning processes. Note that we reproduced this finding using the same model in the Gagne et al. (2020) dataset, so this difference is not introduced by replacing the Bayesian estimation with the MAP parameter estimation. The good quantitative performance of the MOS model and the qualitative explanation of the adaptation effect without introducing a flexible learning rate seem to challenge a range of previous results.

We emphasize that previous results and ours may not be mutually exclusive and may coexist. We argue that the current experimental paradigm is insufficient to disassociate the two possible accounts. Although the MOS framework quantitatively wins in model comparisons, it requires examining their qualitatively distinct predictions to further differentiate the two accounts. We will discuss this issue in future directions below.

The normative interpretation of the mixed strategies

The normative interpretation of learning rate can be elusive. On one hand, the quality of a learning rate does not monotonically increase with its value. Consequently, one's cognitive ability cannot be directly assessed based on their fitted learning rate, unless compared to the theoretically optimal learning rate. On the other hand, the optimal learning rate is highly context-dependent and can even vary from trial to trial (Behrens et al., 2007). This can pose challenges when assessing participants' performance across different cognitive tasks.

Based on the principle of resource rationality, the MOS model demonstrates stronger normative characteristics. The model suggests that preference towards the three strategies can be used to qualitatively approximate this reward-effort tradeoff. Especially, the EU strategy is (defined to be)

the most rewarding strategy (Von Neumann & Morgenstern, 1947), but also cognitively demanding (Gershman et al., 2015). Hence, a higher preference for the EU strategy typically signifies better cognitive ability and capacity. Individuals with psychiatric diseases exhibit a significantly lower preference for this EU strategy compared with healthy individuals, which implies that their cognitive resources might be disrupted. The MO and HA strategies are more computationally economical, though they yield fewer rewards. It is worth noting that the patient group exhibits a greater preference for the MO strategy, which may imply mental impairments beyond limited cognitive resources. According to the resource-rationality principle (Gershman, 2020) and **Fig. 4C**: The HA strategy is a cost-efficient strategy that brings more rewards than the MO strategy. This may prompt further investigation into the underlying reasons behind these preferences.

This framework can be extended to understand human behaviors in paradigms beyond the volatile reversal task. The key lies in identifying heuristics that contrast with the EU strategy. For instance, when employing the MOS model in a volatile reversal task with fixed reward magnitude signals set to 1, we can exclude the MO strategy from the pool and preserve EU and HA. A higher preference for the EU strategy still implies better cognitive ability.

Atypical learning speed in psychiatric diseases

In the present work, we found that patients with depression and anxiety display slower learning speed in the probabilistic learning tasks (shown in **Fig. 4A**). We attributed the very observation to participants' decision preferences. However, in conventional Rescorla-Wagner modeling, the learning speed is primarily indicated by the parameter of learning rate. For example, Chen (2015) conducted a systematic review of reinforcement learning in patients with depression and identified 10 out of 11 behavioral datasets showing either comparable or slower learning rates in depressive patients. Nonetheless, depressive patients may not always have a slower learning rate. In a recent meta-analysis that summarized 27 articles with 3085 participants, including 1242 with depression and/or anxiety, Pike and Robinson (2022) found a reduced reward but enhanced punishment learning rate. This finding yields two practical implications. First, the heterogeneous findings in the literature may arise from heterogeneous pathologies in depression and anxiety. Second, the learning rate as an indicator of human learning and decision-making is not yet perfect and needs to be revised. The mixture decision strategy model may provide useful complementary explanations of the consequences of a spectrum of symptoms.

Limitations and future directions

The MOS model provides relative context-free interpretations for some learning rate phenomena, but not all of them. One among them is the value-specific learning rate differences, where the learning rates for positive outcomes are higher than the negative ones (Chen et al., 2015; Gagne et al., 2020; Pike & Robinson, 2022). It's worth noting that there is no difference between value-specific learning rates, even in MOS 22, where value-specific learning rates are incorporated (Supplementary Note 2). This suggests that at least the effect of the value-specific learning rate is modest in this dataset. Future studies may consider exploring explicit behavioral markers for value-specific learning that do not rely on specific computational models, rather than merely estimating the learning rate value from noisy behavioral data.

We propose an experimental paradigm that can potentially disassociate the learning rate from the mixture of strategy at behavioral level. The idea is to verify the cognitive constraints hypothesis by manipulating participants' cognitive loads. The volatile reversal learning task can be implemented with a secondary task (e.g., asking participants to remember words through headphones). We expect to identify a preference shift from the EU strategy to the MO strategy from participants (a decreasing w_{EU} and an increasing w_{MO}) because human agents should compromise to a simpler

and irrational strategy due to resource constraints induced by the secondary task. In general, we expect this line of research to include more experimental paradigms such that we can gain a complete picture of human learning behavior.

We will also explore why individuals with mental disorders prefer simpler strategies when making decisions. One possible explanation is that individuals with depression exhibit a maladaptive emotion regulation behavior called *rumination*, suffering from irresistible and persistent negative thoughts (Song et al., 2022; Yan et al., 2022). It is likely that the presence of negative thoughts consumes some cognitive resources, such that the participants fail to utilize the complicated but rewarding EU strategy.

Acknowledgements

We thank the authors of Gagne et al. (2020) for sharing their data. This work was supported by the National Natural Science Foundation of China (32100901), Shanghai Pujiang Program (21PJ1407800), Natural Science Foundation of Shanghai (21ZR1434700), the Research Project of Shanghai Science and Technology Commission (20dz2260300) and the Fundamental Research Funds for the Central Universities (to R.-Y.Z.)

Conflict of Interests

The authors declare no competing financial interests.

Author Contributions

Z. F. and R-Y.Z. conceived and designed the study. Z. F., M. Z., T.X., Y.L., H.X., and P.Q. processed the data. Z. F. implemented the computational models. Z. F., and R-Y.Z. made the first draft of the manuscript. All authors provide valuable feedback on the final manuscript.

Supplemental Information

Supplemental Note 1: the priors for reparametrized parameters

We fit the models using the BFGS method, which requires us to first turn the constrained optimization problem (in terms of the parameter range) into an unconstrained one. To do so, we applied the reparameterization tricks. For example, we passed the raw parameter values through the sigmoid function $\xi = \frac{1}{1+\exp(-\xi_{\mathrm{raw}})}$ to create parameters with range (0, 1). For parameters with range (0, ∞), we used the exponential function $\xi = \exp(\xi_{\mathrm{raw}})$. The raw parameter values are all sampled from a Gaussian space.

We carefully tuned the raw parameter priors to ensure the parameters have a reasonable prior or a prior that is consistent with other published research in the reparametrized space (not the raw value space) (**Fig. S1**). For parameters with range (0, 1), we approximate the uniform distribution Uniform(0, 1); for parameters with range (0, ∞), we approximate Gamma(3, 3).

Supplemental Note 2: the complete statistical results of MOS18

We performed multiple 2 × 2 × 2 ANOVAs with the logit of the three weight parameters, dubbed decision preferences, as the dependent variable, group (health control/patients) as a between-subject factor, volatility level (stable/volatile) and feedback types (reward/aversive) as within-subject factors.

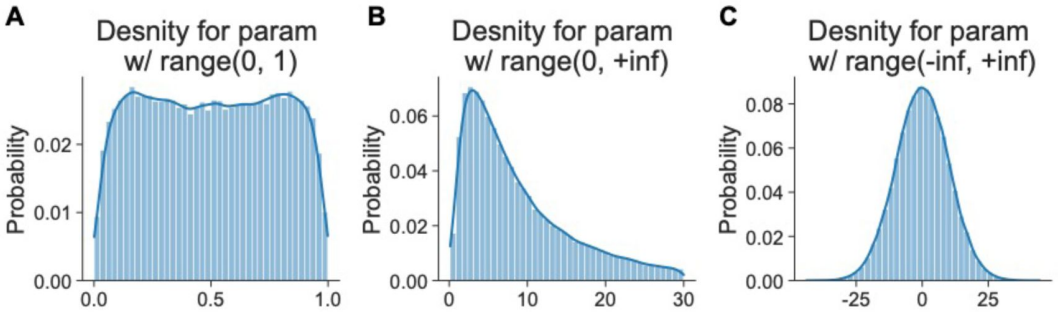


Figure S1

The reparametrized priors for parameters.

A. For parameters with a range of (0, 1), the raw values were sampled from $N(0, 1.55)$ and passed through the sigmoid function.

B. For parameters with a range of (0, ∞), the raw values were sampled from $N(2, 1)$ and passed through the exponential function.

C. For parameters with a range (−∞, ∞), the raw values were sampled from $N(0, 10)$.

For the **weighting parameters for the EU strategy w_{EU}**, the patient group exhibited a weaker tendency for the rational EU strategy ($F(1, 300) = 27.195$, $p < 0.001$, $\eta^2 = 0.076$). People also showed a higher tendency for the EU strategy for the reward feedback than the aversive one ($F(1, 300) = 5.368$, $p = 0.021$, $\eta^2 = 0.016$). No significant main effects of volatility level ($F(1, 300) = 0.022$, $p = 0.926$, $\eta^2 = 0.006$) and significant interaction effects were found (all ps > 0.149).

For the **weighting parameters for the MO strategy w_{MO}**, the patient group exhibited a stronger tendency for the MO strategy ($F(1, 300) = 10.652$, $p < 0.001$, $\eta^2 = 0.031$). No significant main effects of volatility level ($F(1, 300) = 0.537$, $p = 0.464$, $\eta^2 < 0.001$) and feedback types ($F(1, 300) = 0.431$, $p = 0.512$, $\eta^2 = 0.002$) as well as significant interaction effects were found (all ps > 0.420).

For the **weighting parameters for the HA strategy w_{HA}**, the two groups exhibited no significant differences in the preference for the HA strategy ($F(1, 300) = 0.434$, $p = 0.511$, $\eta^2 = 0.001$). No significant main effects of volatility level ($F(1, 300) = 0.872$, $p = 0.351$, $\eta^2 = 0.003$) and feedback types ($F(1, 300) = 1.484$, $p = 0.224$, $\eta^2 = 0.004$) as well as significant interaction effects were found (all ps > 0.357).

For the **log learning rates log α_ψ**, there were no significant main effects of participant groups ($F(1, 300) = 1.489$, $p = 0.223$, $\eta^2 = 0.002$), feedback types ($F(1, 300) = 0.002$, $p = 0.961$, $\eta^2 = 0.000$), and volatility levels ($F(1, 300) = 1.280$, $p = 0.258$, $\eta^2 = 0.002$). We also examined the value-specific learning rate effect and not significant difference ($F(1, 300) = 0.006$, $p = 0.937$, $\eta^2 = 0.000$). There is a weak significance in the interaction patients group × volatility levels × feedback types ($F(1, 300) = 3.998$, $p = 0.046$, $\eta^2 = 0.006$). No other significant interaction effects were found (all ps > 0.258).

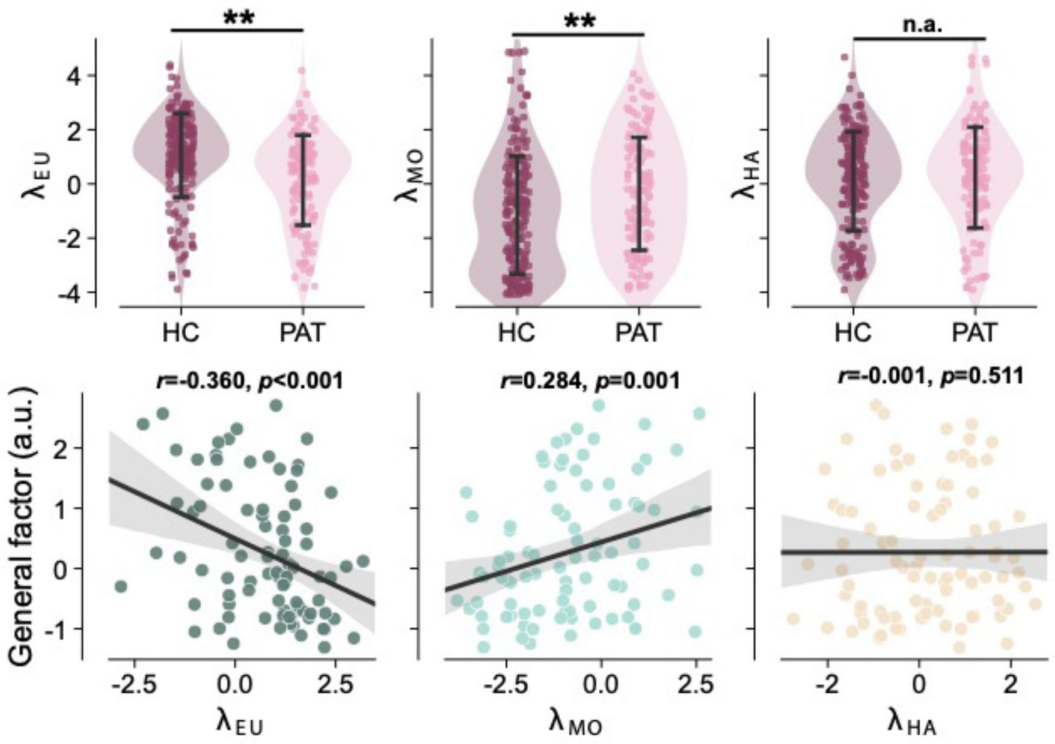


Figure S2

Parameter analyses of MOS22.

A. The weighting parameters for the healthy control participants (HC) and the patients (PAT) diagnosed with MDD and GAD. The y-axis means averaged preference over different volatility levels (volatile and stable) and feedback types (reward and aversive). Error bars reflect the standard error of the mean value across participants ×experimental conditions× feedback types.

B. Decision preferences predict participants' general factor score (*g* score) in the bifactor analysis reported in Gagne, et al. (2020). The y-axis indicates the averaged preference over different volatility levels (volatile and stable) and feedback types (reward and aversive). This average operation is permitted here because the logit of the weight is normally distributed. The shaded areas reflect 95% confidence intervals of the regression prediction.

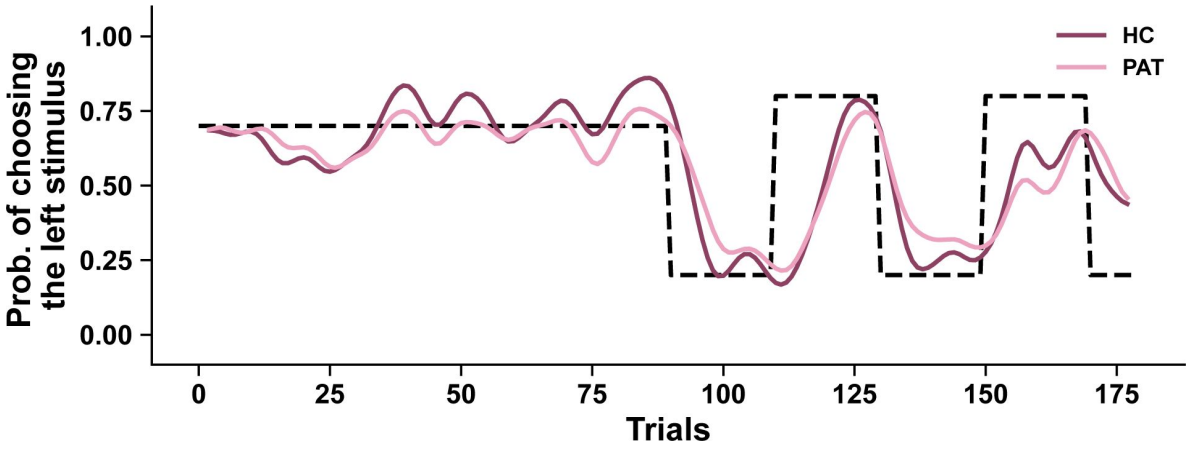


Figure S3

The human learning behaviors of the two groups in the reward condition.

References

Akaike H. (1974) **A new look at the statistical model identification** *IEEE transactions on automatic control* **19**:716–723

Beck A. T., Ward C., Mendelson M., Mock J., Erbaugh J. (1961) **Beck depression inventory (BDI)** *Arch Gen Psychiatry* **4**:561–571

Behrens T. E., Woolrich M. W., Walton M. E., Rushworth M. F. (2007) **Learning the value of information in an uncertain world** *Nat Neurosci* **10**:1214–1221

Boyd S. P., Vandenberghe L. (2004) **Convex optimization**

Browning M., Behrens T. E., Jocham G., O'Reilly J. X., Bishop S. J. (2015) **Anxious individuals have difficulty learning the causal statistics of aversive environments** *Nat Neurosci* **18**:590–596

Chen C., Takahashi T., Nakagawa S., Inoue T., Kusumi I. (2015) **Reinforcement learning in depression: A review of computational research** *Neuroscience & Biobehavioral Reviews* **55**:247–267

Clark L. A., Watson D. (1991) **Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications** *J Abnorm Psychol* **100**:316–336

Cohen A. S., McGovern J. E., Dinzeo T. J., Covington M. A. (2014) **Cohen, A. S., McGovern, J. E., Dinzeo, T. J., & Covington, M. A. (2014). Speech deficits in serious mental illness: a cognitive resource issue? Schizophr Res, 160(), 173–179.** *Schizophr Res* **160**:173–179

Cools R., Clark L., Owen A. M., Robbins T. W. (2002) **Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging** *J Neurosci* **22**:4563–4567

Daw N. D., Gershman S. J., Seymour B., Dayan P., Dolan R. J. (2011) **Model-based influences on humans' choices and striatal prediction errors** *Neuron* **69**:1204–1215

Delacre M., Lakens D., Leys C. (2017) **Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test** *International Review of Social Psychology* **30**

Eysenck H. J., Eysenck S. B. (1975) **Eysenck personality questionnaire (junior & adult)**

Fan H., Gershman S. J., Phelps E. A. (2023) **Trait somatic anxiety is associated with reduced directed exploration and underestimation of uncertainty** *Nature Human Behaviour* **7**:102–113

Gagne C., Zika O., Dayan P., Bishop S. J. (2020) **Impaired adaptation of learning to contingency volatility in internalizing psychopathology** *Elife* **9**

Gershman S. J. (2020) **Origin of perseveration in the trade-off between reward and complexity** *Cognition* **204**

Gershman S. J., Horvitz E. J., Tenenbaum J. B. (2015) **Computational rationality: A converging paradigm for intelligence in brains, minds, and machines** *Science* **349**:273–278

Gilovich T., Vallone R., Tversky A. (1985) **The hot hand in basketball: On the misperception of random sequences** *Cognitive psychology* **17**:295–314

Griffiths T. L., Lieder F., Goodman N. D. (2015) **Rational use of cognitive resources: levels of analysis between the computational and the algorithmic** *Top Cogn Sci* **7**:217–229

Harvey P. O., Fossati P., Pochon J. B., Levy R., Lebastard G., Lehericy S., Allilaire J. F., Dubois B. (2005) **Cognitive control and brain resources in major depression: an fMRI study using the n-back task** *Neuroimage* **26**:860–869

Lawson R. P., Mathys C., Rees G. (2017) **Adults with autism overestimate the volatility of the sensory environment** *Nat Neurosci* **20**:1293–1299

Levens S. M., Muhtadie L., Gotlib I. H. (2009) **Rumination and impaired resource allocation in depression** *J Abnorm Psychol* **118**:757–766

Meyer T. J., Miller M. L., Metzger R. L., Borkovec T. D. (1990) **Development and validation of the Penn State Worry Questionnaire** *Behav Res Ther* **28**:487–495

Moran T. P. (2016) **Anxiety and working memory capacity: A meta-analysis and narrative review** *Psychol Bull* **142**:831–864

Pearce J. M., Hall G. (1980) **A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli** *Psychological review* **87**

Pike A. C., Robinson O. J. (2022) **Reinforcement Learning in Patients With Mood and Anxiety Disorders vs Control Individuals: A Systematic Review and Meta-analysis** *JAMA psychiatry* **79**:313–322

Powers A. R., Mathys C., Corlett P. R. (2017) **Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors** *Science* **357**:596–600

Radloff L. S. (2016) **The CES-D Scale** *Applied psychological measurement* **1**:385–401

Rescorla R. A. (1972) **A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement** *Current research and theory* :64–99

Rigoux L., Stephan K. E., Friston K. J., Daunizeau J. (2014) **Bayesian model selection for group studies—revisited** *Neuroimage* **84**:971–985

Schwarz G. (1978) **Estimating the dimension of a model** *The annals of statistics* :461–464

Song X., Long J., Wang C., Zhang R., Lee T. M. (2022) **The inter-relationships of the neural basis of rumination and inhibitory control: neuroimaging-based meta-analyses** *Psychoradiology* **2**:11–22

Spielberger CD G. R., Lushene R, Vagg PR, Jacobs GA (1983) **Manual for the State-Trait Anxiety Inventory** *Consulting Psychologists Press*

Sutton R. S., Barto A. G. (2018) **Reinforcement learning: An introduction**

Von Neumann J., Morgenstern O. (1947) **Von Neumann, J., & Morgenstern, O. (1947). Theory of games and economic behavior, 2nd rev.** *Theory of games and economic behavior*

Watson D., Clark L. A. (1991) **Mood and anxiety symptom questionnaire** *Journal of Behavior Therapy and Experimental Psychiatry*

Wood W., Runger D. (2016) **Psychology of Habit** *Annu Rev Psychol* **67**:289–314

Yan X., Gao W., Yang J., Yuan J. (2022) **Emotion regulation choice in internet addiction: less reappraisal, lower frontal alpha asymmetry** *Clinical EEG and Neuroscience* **53**:278–286

Article and author information

Zeming Fang
Shanghai Mental Health, School of Medicine, Shanghai Jiao Tong University, Shanghai, 200030, China, Institute of Psychology and Behavioral Science, Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, 200030, China
ORCID iD: 0000-0002-8091-4413

Meihua Zhao
School of Psychology, South China Normal University, Guangzhou, 510631, China
ORCID iD: 0000-0002-2238-0513

Ting Xu
The Center of Psychosomatic Medicine, Sichuan Provincial Center for Mental Health, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, 611731, China
ORCID iD: 0000-0003-1590-9025

Yuhang Li
Centre of Centre for Cognitive and Brain Sciences, Institute of Collaborative Innovation, University of Macau, Macau, 999078, China

Hanbo Xie
Department of Psychology, University of Arizona, Tucson, Arizona, USA
ORCID iD: 0000-0003-1133-8544

Peng Quan
Research Center for Quality of Life and Applied Psychology, Guangdong Medical University, Dongguan, China
ORCID iD: 0000-0001-5416-536X

Haiyang Geng
Tianqiao and Chrissy Chen Institute for Translational Research, Shanghai, 200040, China
ORCID iD: 0000-0001-6115-807X

Ru-Yuan Zhang
Shanghai Mental Health, School of Medicine, Shanghai Jiao Tong University, Shanghai, 200030, China, Institute of Psychology and Behavioral Science, Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, 200030, China, Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, School of Psychology and Cognitive Science, East China Normal University, Shanghai, 200241, China
For correspondence: ruyuanzhang@sjtu.edu.cn
ORCID iD: 0000-0002-0654-715X

Copyright

Editors

Reviewing Editor
Mimi Liljeholm
University of California, Irvine, Irvine, United States of America

Senior Editor
Michael Frank
Brown University, Providence, United States of America

Reviewer #1 (Public Review):

Summary:
This paper describes a reanalysis of data collected by Gagne et al. (2020), who investigated how human choice behaviour differs in response to changes in environmental volatility. Several studies to date have demonstrated that individuals appear to increase their learning rate in response to greater volatility and that this adjustment is reduced amongst individuals with anxiety and depression. The present authors challenge this view and instead describe a novel Mixture of Strategies (MOS) model, that attributes individual differences in choice behaviour to different weightings of three distinct decision-making strategies. They demonstrate that the MOS model provides a superior fit to the data and that the previously observed differences between patients and healthy controls may be explained by patients opting for a less cognitively demanding, but suboptimal, strategy.

Strengths:
The authors compare several models (including the original winning model in Gagne et al., 2020) that could feasibly fit the data. These are clearly described and are evaluated using a range of model diagnostics. The proposed MOS model appears to provide a superior fit across several tests.

The MOS model output is easy to interpret and has good face validity. This allows for the generation of clear, testable, hypotheses, and the authors have suggested several lines of potential research based on this.

Weaknesses:
The authors justify this reanalysis by arguing that learning rate adjustment (which has previously been used to explain choice behaviour on volatility tasks) is likely to be too

computationally expensive and therefore unfeasible. It is unclear how to determine how "expensive" learning rate adjustment is, and how this compares to the proposed MOS model (which also includes learning rate parameters), which combines estimates across three distinct decision-making strategies.

As highlighted by the authors, the model is limited in its explanation of previously observed learning differences based on outcome value. It's currently unclear why there would be a change in learning across positive/negative outcome contexts, based on strategy choice alone.

https://doi.org/10.7554/eLife.93887.1.sa2

Reviewer #2 (Public Review):

Summary:
Previous research shows that humans tend to adjust learning in environments where stimulus-outcome contingencies become more volatile. This learning rate adaptation is impaired in some psychiatric disorders, such as depression and anxiety. In this study, the authors reanalyze previously published data on a reversal-learning task with two volatility levels. Through a new model, they provide some evidence for an alternative explanation whereby the learning rate adaptation is driven by different decision-making strategies and not learning deficits. In particular, they propose that adjusting learning can be explained by deviations from the optimal decision-making strategy (based on maximizing expected utility) due to response stickiness or focus on reward magnitude. Furthermore, a factor related to the general psychopathology of individuals with anxiety and depression negatively correlated with the weight on the optimal strategy and response stickiness, while it correlated positively with the magnitude strategy (a strategy that ignores the probability of outcome).

Strengths:
The main strength of the study is a novel and interesting explanation of an otherwise well-established finding in human reinforcement learning. This proposal is supported by rigorously conducted parameter retrieval and the comparison of the novel model to a wide range of previously published models.

Weaknesses:
My main concern is that the winning model (MOS6) does not have an error term (inverse temperature parameter beta is fixed to 8.804).

1. It is not clear why the beta is not estimated and how were the values presented here chosen. It is reported as being an average value but it is not clear from which parameter estimation. Furthermore, with an average value for participants that would have lower values of inverse temperature (more stochastic behaviour) the model is likely overfitting.

2. In the absence of a noise parameter, the model will have to classify behaviour that is not explained by the optimal strategy (where participants simply did not pay attention or were not motivated) as being due to one of the other two strategies.

3. A model comparison among models with inverse temperature and variable subsets of the three strategies (EU + MO, EU + HA) would be interesting to see. Similarly, comparison of the MOS6 model to other models where the inverse temperature parameter is fixed to 8.804).

This is an important limitation because the same simulation as with the MOS model in Figure 3b can be achieved by a more parsimonious (but less interesting) manipulation of the inverse temperature parameter.

Furthermore, the claim that the EU represents an optimal strategy is a bit overstated. The EU strategy is the only one of the three that assumes participants learn about the stimulus-outcomes contingencies. Higher EU strategy utilisation will include participants that are more optimal (in maximum utility maximisation terms), but also those that just learned better and completely ignored the reward magnitude.

Other minor issues that I have are the following:
The mixture strategies model is an interesting proposal, but seems to be a very convoluted way to ask: to what degree are decisions of subjects affected by reward, what they've learned, and response stickiness? It seems to me that the same set of questions could be addressed with a simpler model that would define choice decisions through a softmax with a linear combination of the difference in rewards, the difference in probabilities, and a stickiness parameter.

Learning rate adaptation was also shown with tasks where decision-making strategies play a less important role, such as the Predictive Inference task (see for instance Nassar et al, 2010). When discussing the merit of the findings of this study on learning rate adaptation across volatility blocks, this work would be essential to mention.

https://doi.org/10.7554/eLife.93887.1.sa1

Reviewer #3 (Public Review):

Summary:
This paper presents a new formulation of a computational model of adaptive learning amid environmental volatility. Using a behavioral paradigm and data set made available by the authors of an earlier publication (Gagne et al., 2020), the new model is found to fit the data well. The model's structure consists of three weighted controllers that influence decisions on the basis of (1) expected utility, (2) potential outcome magnitude, and (3) habit. The model offers an interpretation of psychopathology-related individual differences in decision-making behavior in terms of differences in the relative weighting of the three controllers.

Strengths:
The newly proposed "mixture of strategies" (MOS) model is evaluated relative to the model presented in the original paper by Gagne et al., 2020 (here called the "flexible learning rate" or FLR model) and two other models. Appropriate and sophisticated methods are used for developing, parameterizing, fitting, and assessing the MOS model, and the MOS model performs well on multiple goodness-of-fit indices. The parameters of the model show decent recoverability and offer a novel interpretation for psychopathology-related individual differences. Most remarkably, the model seems to be able to account for apparent differences in behavioral learning rates between high-volatility and low-volatility conditions even with no true condition-dependent change in the parameters of its learning/decision processes. This finding calls into question a class of existing models that attribute behavioral adaptation to adaptive learning rates.

Weaknesses:
1. Some aspects of the paper, especially in the methods section, lacked clarity or seemed to assume context that had not been presented. I found it necessary to set the paper down and read Gagne et al., 2020 in order to understand it properly.

2. There is little examination of why the MOS model does so well in terms of model fit indices. What features of the data is it doing a better job of capturing? One thing that makes this puzzling is that the MOS and FLR models seem to have most of the same qualitative components: the FLR model has parameters for additive weighting of magnitude relative to

probability (akin to the MOS model's magnitude-only strategy weight) and for an autocorrelative choice kernel (akin to the MOS model's habit strategy weight). So it's not self-evident where the MOS model's advantage is coming from.

3. One of the paper's potentially most noteworthy findings (Figure 5) is that when the FLR model is fit to synthetic data generated by the expected utility (EU) controller with a fixed learning rate, it recovers a spurious difference in learning rate between the volatile and stable environments. Although this is potentially a significant finding, its interpretation seems uncertain for several reasons:

- According to the relevant methods text, the result is based on a simulation of only 5 task blocks for each strategy. It would be better to repeat the simulation and recovery multiple times so that a confidence interval or error bar can be estimated and added to the figure.

- It makes sense that learning rates recovered for the magnitude-oriented (MO) strategy are near zero, since behavior simulated by that strategy would have no reason to show any evidence of learning. But this makes it perplexing why the MO learning rate in the volatile condition is slightly positive and slightly greater than in the stable condition.

- The pure-EU and pure-MO strategies are interpreted as being analogous to the healthy control group and the patient group, respectively. However, the actual difference in estimated EU/MO weighting between the two participant groups was much more moderate. It's unclear whether the same result would be obtained for a more empirically plausible difference in EU/MO weighting.

- The fits of the FLR model to the simulated data "controlled all parameters except for the learning rate parameters across the two strategies" (line 522). If this means that no parameters except learning rate were allowed to differ between the fits to the pure-EU and pure-MO synthetic data sets, the models would have been prevented from fitting the difference in terms of the relative weighting of probability and magnitude, which better corresponds to the true difference between the two strategies. This could have interfered with the estimation of other parameters, such as learning rate.

- If, after addressing all of the above, the FLR model really does recover a spurious difference in learning rate between stable and volatile blocks, it would be worth more examination of why this is happening. For example, is it because there are more opportunities to observe learning in those blocks?

4. Figure 4C shows that the habit-only strategy is able to learn and adapt to changing contingencies, and some of the interpretive discussion emphasizes this. (For instance, line 651 says the habit strategy brings more rewards than the MO strategy.) However, the habit strategy doesn't seem to have any mechanism for learning from outcome feedback. It seems unlikely it would perform better than chance if it were the sole driver of behavior. Is it succeeding in this example because it is learning from previous decisions made by the EU strategy, or perhaps from decisions in the empirical data?

5. For the model recovery analysis (line 567), the stated purpose is to rule out the possibility that the MOS model always wins (line 552), but the only result presented is one in which the MOS model wins. To assess whether the MOS and FLR models can be differentiated, it seems necessary also to show model recovery results for synthetic data generated by the FLR model.

6. To the best of my understanding, the MOS model seems to implement valence-specific learning rates in a qualitatively different way from how they were implemented in Gagne et al., 2020, and other previous literature. Line 246 says there were separate learning rates for upward and downward updates to the outcome probability. That's different from using two learning rates for "better"- and "worse"-than-expected outcomes, which will depend on both the direction of the update and the valence of the outcome (reward or shock). Might this

relate to why no evidence for valence-specific learning rates was found even though the original authors found such evidence in the same data set?

7. The discussion (line 649) foregrounds the finding of greater "magnitude-only" weights with greater "general factor" psychopathology scores, concluding it reflects a shift toward simplifying heuristics. However, the picture might not be so straightforward because "habit" weights, which also reflect a simplifying heuristic, correlated negatively with the psychopathology scores.

https://doi.org/10.7554/eLife.93887.1.sa0