communications psychology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s44271-025-00249-y

Distributional dual-process model predicts strategic shifts in decision-making under uncertainty

Mianzhi Hu [1] ✉, Hilary J. Don[1,2] & Darrell A. Worthy[1]

In an uncertain world, human decision-making often involves adaptively leveraging different strategies to maximize gains. These strategic shifts, however, are overlooked by many traditional reinforcement learning models. Here, we incorporate parallel evaluation systems into distribution-based modeling and propose an entropy-weighted dual-process model that leverages Dirichlet and multivariate Gaussian distributions to represent frequency and value-based decision-making strategies, respectively. Model simulations and empirical tests demonstrated that our model outperformed traditional RL models by uniquely capturing participants' strategic change from value-based to frequency-based learning in response to heightened uncertainty. As reward variance increased, participants switched from focusing on actual rewards to using reward frequency as a proxy for value, thereby showing greater preference for more frequently rewarded but less valuable options. These findings suggest that increased uncertainty encourages the compensatory use of diverse evaluation methods, and our dual-process model provides a promising framework for studying multi-system decision-making in complex, multivariable contexts.

Human decision-making has been extensively studied through the lens of expected value (EV) or expected utility (EU), which quantitatively defines the anticipated outcome of choosing a given option, adjusted by subjective weightings for, for example, outcome valence, probability, and magnitude[1–3]. Since the 17th century, EV and EU typically reflect holistic, unitary, and deterministic measures of utility, which can be rational, or, more often, biased under conditions of uncertainty and/or incomplete information[4–7]. Researchers have long sought to develop computational models that can accurately specify the cognitive pathways people use to adjust their expectations, usually in the context of repeated reinforcement learning (RL) and under constraints of limited time and resources[8–10]. Yet, the applicability and effectiveness of many current RL models still vary significantly across different contexts[8,9,11], suggesting that EV may not be best represented as a singular value[12–14], nor do individuals consistently adhere to one single decision-making strategy[15–17].

Emerging evidence from neural and behavioral modeling studies suggests that, rather than aiming for a definite EV, people are more likely to process incoming information in a probabilistic manner, continuously updating their internal representations of expected outcomes as distributions[12,18–20]. Building upon this, we propose that distribution-based modeling with parallel evaluation systems can be leveraged to address the limitations of both singular EV representations and adaptive strategy switching. In this framework, each system represents expected outcomes as distinct distributions, which are updated independently using only relevant aspects of the outcome information and weighted by their associated statistical dispersion (i.e., entropy) to inform final decisions. In this work, we demonstrate that an entropy-weighted dual-process model that integrates two simple distribution functions—namely the Dirichlet and the multivariate Gaussian distribution—can effectively capture frequency-based and value-based decision-making strategies, adapt to diverse RL contexts, and elucidate the underlying rationale behind trial-by-trial decisions.

In instance-based RL, a critical finding from previous research on adaptive decision-making reveals that decisions are guided not only by the EV of rewards but also by their frequencies[4,9–11,21,22]. Reward frequency sometimes serves as a proxy for the perceived value of an option, giving people an intuitive sense of which option provides more rewards. When making decisions based on experience, people tend to undervalue infrequently rewarded, less familiar options, often leading to a disproportional preference for more frequently rewarded options even when they yield suboptimal long-term payoffs[8,11,16,23]. Interestingly, these frequency effects manifest only between pairs of options with marginal value differences, while in cases where the difference is more discernible, the frequency of reinforcement has little impact on decision-making[11,23].

Supporting this, cross-model comparison studies[9,11] also show that frequency-sensitive models (e.g., Decay models) better fit participants' behavior than mean-centered models (e.g., the Delta model) in more

[1]Texas A&M University, College Station, TX, USA. [2]University College London, London, UK. ✉e-mail: rudolfhu@tamu.edu

information-intensive, memory-demanding scenarios, such as four-option tasks or an omniscient version of the Iowa Gambling Task (IGT) where participants have full information about the outcomes of all decks regardless of their choices. However, this advantage dissipates in simpler tasks, such as binary choice tasks and the original IGT. These findings point to a potential multivariable interplay among reward value, frequency, and self-perceived uncertainty, where individuals tend to make rational, value-based decisions when they are able to fully process option values, whereas under conditions of high value uncertainty, they may consult a more intuitive, reflexive decision-making system that prioritizes the frequency of rewards as a proxy for value, leading to habitual behavioral patterns favoring the safe option.

Unfortunately, many existing RL models fail to account for this uncertainty-driven strategic shift, as they either computationally overlook the role of uncertainty, or lack the flexibility to accommodate alternative decision-making strategies. Among those models that do incorporate uncertainty, asymmetric RL models[24] that assign different learning rates for positive versus negative prediction errors have been shown to account for risk sensitivity in both neural and behavioral data. Similarly, the mean-variance utility model from the economics literature also discounts an option's EV by its estimated variance[25]. However, as our simulation results later demonstrate, these models do not predict a shift in preference from more valuable options under low value uncertainty to more frequently rewarded options under high value uncertainty. This is because these models apply the same level of discounting to the EVs of all options when reward variance is uniformly elevated, leaving no mechanism to differentiate preferences. Without a supplementary system to aid in the decision-making process under such conditions, these uncertainty-sensitive models fail to capture frequency effects in environments where variance is equally high across all options.

To empirically test our model and explore these nuanced interactions, we presented participants with four options (A, B, C, D) across three randomly assigned levels of reward variance: low variance (LV), moderate variance (MV), and high variance (HV). Rewards for each option were randomly drawn from a normal distribution with constant mean values (A = 0.65, B = 0.35, C = 0.75, D = 0.25) and the assigned variance level. The task consisted of two phases[11]. In the first phase, participants underwent 150 training trials with feedback, selecting from either AB or CD pairs. In each pair, one option (i.e., A and C) yielded significantly higher rewards than the other (i.e., B and D). However, AB pairs were presented twice as often as CD pairs (i.e., 100 AB trials versus 50 CD trials), resulting in A providing more frequent rewards than C, despite having a lower average reward value (A: 0.65, C: 0.75). After the training phase, participants proceeded to the transfer phase, where they selected from the remaining combinations (i.e., AD, BD, CA, CB) without feedback. We were particularly interested in examining selections from the new CA pair to gauge participants' subjective evaluation of C versus A, as this pair presents a scenario where the more frequently rewarded option (A) has a lower average payoff than the other option (C).

We hypothesized that (1) EV in multivariable scenarios is better represented as an entropy-weighted sum of estimations from parallel distributional evaluative systems than a singular value generated by a fixed learning rule; (2) in our task, people's preference for the more frequent but less rewarding option, A, would increase with higher reward variance due to increased entropy—or reduced confidence—in value-based processing and the need to explore alternative evaluation strategies; (3) this shift in preference would be correlated with increased model-inferred weights on frequency-based processing and a corresponding decreased reliance in value-based processing.

Methods

Participants

This study was approved by Texas A&M University Institutional Review Board (IRB2021-1008M). Participants were 293 undergraduate students from Texas A&M University, who received partial course credit for their participation and provided informed consent. The mean age was 19.13 years (SD = 1.58), with 65% of participants self-identifying as women (Women = 169, Men = 91, Missing = 33). Demographic information for 33 of the 293 participants was not collected due to technical issues. However, as most participants were first-year students enrolled in the introductory psychology course, the overall distribution of age and sex is very unlikely to be significantly impacted by the missing data. Race and ethnicity information was not collected but is expected to align with the typical demographic profile of first-year psychology students at Texas A&M University. Participants were randomly allocated into the LV (n = 93), MV (n = 100), and HV (n = 100) conditions. No participant was excluded from our analysis.

Procedures

The experiment was run on Windows lab computers using Psychtoolbox 3 for MATLAB. The procedures are fundamentally the same as Don et al. (2019). Figure 1 shows an example trial sequence of the task. Participants were instructed that they would make repeated choices on each trial, and they would gain or lose rewards for each choice. They were told their goal was to gain as many points as possible, so they should try to learn which options are most rewarding. Choice options were four fractal images, presented in different pair combinations. The fractals each participant saw were randomly drawn from a pool of 12 images and the order of the 4 selected images were randomly arranged on screen. Further, the AB and CD option pairs' placement was also randomized onscreen. AB and CD were always together as a pair, but the order of each pair varied for each participant. As an example, some potential orderings of the option could include: ABCD, CDAB, BACD, etc. The reward structure is shown in Table 1.

In the training phase, participants made choices between option A versus option B, or option C vs option D. There were 100 AB trials, and 50 CD trials, with each trial type randomly distributed over the 150 training trials. Rewards were centered at each deck's mean value (A = 0.65, B = 0.35, C = 0.75, D = 0.25). For the HV condition, rewards were normally distributed with the standard deviation mirroring that of a binomial distribution. For decks A and B, this was $(.65 \times .35)^{.5} = .48$, and for C and D this was $(.75 \times .25)^{.5} = .43$. For the MV condition, the standard deviation in rewards was half that of the HV condition (i.e., .24 for AB; .22 for CD), and the LV condition was a quarter of those in the HV condition (i.e., 0.12 for AB; .11 for CD)

After training, participants were told they would now choose between different option pairs. Transfer test trials included 25 of each AC, AD, BC and BD trials, in random order. No feedback was provided for choices in the test phase.

This study was not preregistered; however, our experiment was carefully designed and provided clear predictions regarding the expected direction of the effects. Data distribution was assumed to be normal but this was not formally tested.

Model

In our model, we innovatively leverage two widely used distribution functions—the Dirichlet and multivariate Gaussian distributions—to represent frequency-based and value-based decision-making processes, respectively, thereby minimizing presumptions and manual engineering[26]. The multivariate Gaussian distribution, akin to a univariate normal distribution, defines the EV distribution of each option solely based on the reward's mean and variance, without retaining a tally of past rewards. Previous models have included additional parameters to account for aspects of subjective utility such as loss aversion and discounting of large rewards[3], but in this work, since most rewards were gains, and rewards were normally distributed, we assumed that rewards were processed veridically, rather than as subjective utility. As a result, we focused on testing simpler models that do not make assumptions regarding subjective utility. Additionally, as Yechiam (2019) pointed out, EU is often linear and symmetrical when the absolute magnitude of EVs is relatively small (e.g., below 100), making a subjective utility function unnecessary in such cases[27]. In our paradigm, the magnitude of outcomes ranged between 0 and 1, which falls well within this range. During model comparison,

however, we included one model—the mean-variance utility model—which incorporates a shaping parameter for the utility function.

In contrast, the Dirichlet distribution, as a multivariate extension of the Beta distribution, records only the number of successes and is commonly used as an a priori distribution to estimate the probabilities of multiple events with binary outcomes (i.e., successes and failures)[28,29]. In our model, we use the Dirichlet distribution to track the success frequencies across the four options, essentially recording how many times each option has been chosen and yielded a rewarding outcome. Given that our RL paradigm involves mostly gains, we define a "success" as receiving a reward higher than the average of the estimated mean values across all four options in the Gaussian process. On each trial, if the received reward exceeds this overall average, one success is added to the chosen option in the Dirichlet function. This causes the distribution function to allocate slightly more probability density to that option for future selections. Although the defining a "success" in the Dirichlet distribution may initially require input from the Gaussian process, this information is quickly converted into an offline tally of past rewarding instances (i.e., the α parameter in the Dirichlet distribution). During decision-making, the Dirichlet distribution involves only the retrieval of

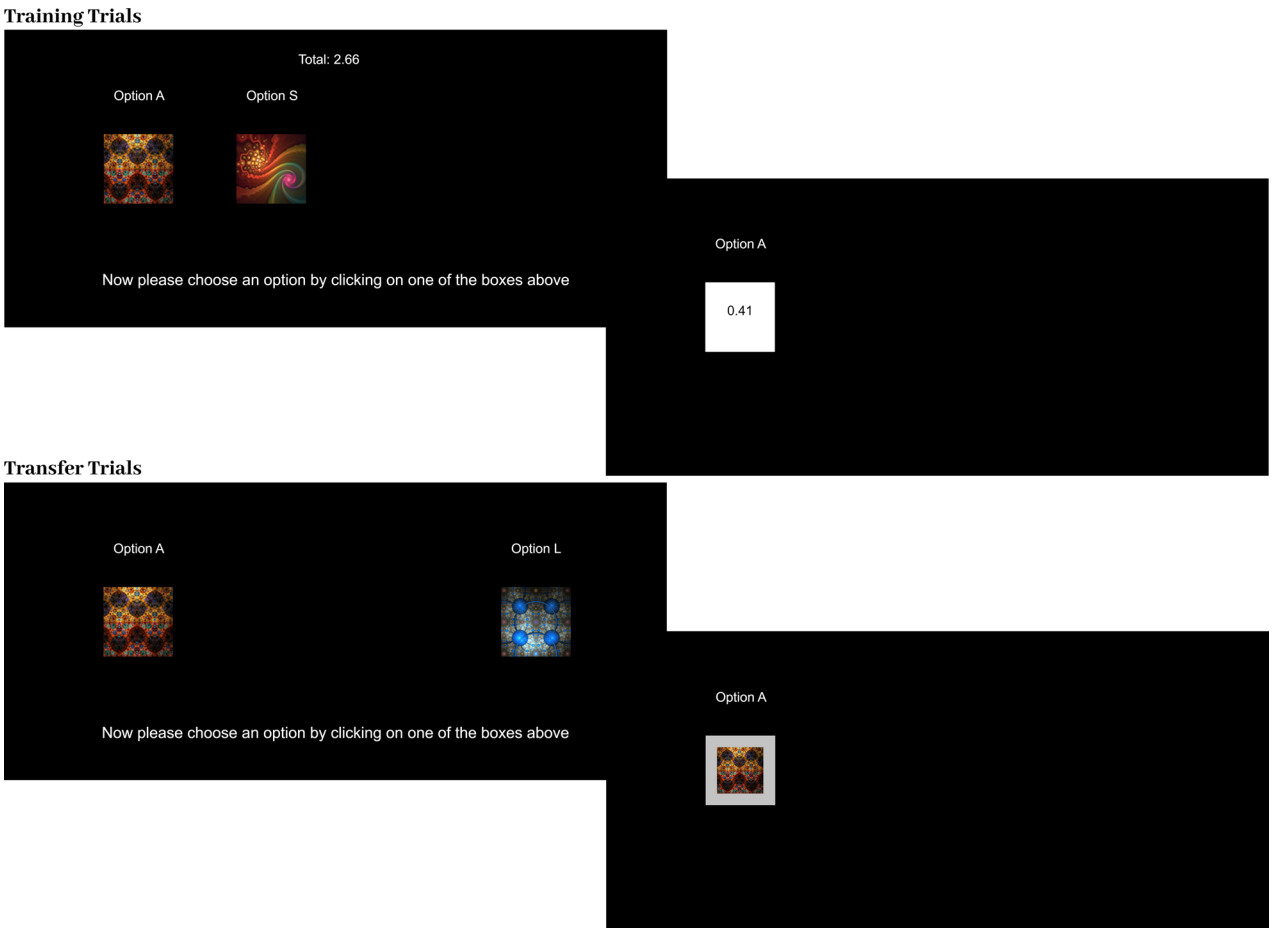


Fig. 1 | Example trial sequences. During the first 150 training trials, participants selected from either the AB or the CD pair. Option labels were randomized (e.g., Option S could correspond to A, B, C, or D in the reward schedule). The total virtual points earned were displayed at the top of the screen. After each selection, participants received feedback on the number of points gained. In the subsequent 100 transfer trials, participants chose from the remaining four option combinations (i.e., 25 trials for each)— CA, BD, AD, and CB—without cumulative point displays or feedback. They were instructed to rely on knowledge acquired during the training phase to maximize their points.

Table 1 | Reward Structure

		Option			
Group		A	B	C	D
Low Variance	N(M,SD)	0.65(0.48)	0.35(0.48)	0.75(0.43)	0.25(0.43)
	base-rate	2		1	
Moderate Variance	N(M,SD)	0.65(0.24)	0.35(0.24)	0.75(0.22)	0.25(0.22)
	base-rate	2		1	
High Variance	N(M,SD)	0.65(0.12)	0.35(0.12)	0.75(0.11)	0.25(0.11)
	base-rate	2		1	

N(M,SD) indicates continuous normal distributions of rewards for each option. M is the mean and SD is the standard deviation. "Base-rate" indicates how frequently each choice pair is presented during training, relative to the other choice pair. For example, 2:1 means the first pair is presented twice as often as the second pair.

this tally, which is hypothesized to be less resource-consuming than the Gaussian process.

These two distribution functions allow us to dissociate the impact of reward frequency from actual reward values. The Dirichlet and multivariate Gaussian distributions share few elements in generating the posterior distribution. The Dirichlet distribution retains only the number of successes per option, while the Gaussian distribution estimates the underlying value distributions. The contrast between these two distributions becomes particularly evident when considering an option that has been rewarded frequently but with small rewards. The Dirichlet distribution would assign a high probability mass to this option because of its high reward frequency, whereas the multivariate Gaussian distribution would show little preference, as the small rewards do not significantly alter the perceived average value for that option (see Fig. 2).

A major challenge with integrating multiple decision-making systems is determining their relative weight. Previous studies juxtaposing two processes often use a constant weight parameter to reflect the overall preference for one process over the other[15,17,30–33]. However, this approach assumes that the same weight is applied to each choice the participant makes, which is likely not the case, especially in cross-comparisons among multiple options. For instance, in our task, people probably do not apply the same weight when choosing between options A and B, where A is clearly better than B, as they would when choosing between options A and C, where the difference is more ambiguous.

To address this, we conceptualize that individuals should give greater weight to the estimation process with less uncertainty. In modeling terms, this implies that the distribution with lower entropy should be given a higher weight. Gaussian entropy reflects confidence in the estimated value of an option (e.g., "How sure am I that option A is worth this much?"), while Dirichlet entropy reflects confidence in the observed frequency of rewards for an option (e.g., "How sure am I that option A has yielded this many rewards so far?"). In this work, the term "uncertainty" refers specifically to value uncertainty in the Gaussian process, influenced by our reward variance manipulation. In contrast, frequency uncertainty, represented by the statistical dispersion of the Dirichlet distribution, will be referred to separately as "frequency". This modeling approach creates a delicate balance between uncertainty and EV, where the impact of EV differences will be proportional to the level of uncertainty within the estimation process. In other words, an estimation process will have a large impact only when the EV difference is substantial, and the individual is confident about the difference. Specifically, uncertainty in a given distribution function $f(x)$ is estimated using differential entropy $h(x)$[34]:

$$h(x) = -\int_{\mathcal{X}} f(x)\log[f(x)]dx \tag{1}$$

and the weight for the Dirichlet process w_D is calculated as:

$$w_D = \frac{w_1 \cdot 2^{h(G)}}{w_1 \cdot 2^{h(G)} + (1 - w_1) \cdot 2^{h(D)}} \tag{2}$$

where $h(G)$ and $h(D)$ represent the differential entropies for the multivariate Gaussian and Dirichlet distributions, respectively (see *Supplementary Methods* for the definition of individual distribution functions). The term $2^{h(x)}$ converts differential entropy into effective volume, which can be interpreted as the size of the space occupied by a continuous random variable x[34]. It also assures that the measure of uncertainty is always positive. The parameter w_1 denotes the subjective preference for the Dirichlet distribution, analogous to a constant weight parameter, which captures people's baseline reliance on the Dirichlet process. Crucially, the overall Dirichlet weight, w_D, is the product of the objective weight—calculated as the proportion of inversed Dirichlet entropy relative to the cumulative inversed entropy—and the subjective weight, which accounts for individual differences. This combined weight is then applied to the probabilities generated from a *SoftMax* rule. The predicted probability that option j will be chosen on trial t, $P|C_j(t)|$ is calculated as:

$$P|C_j(t)| = w_D \cdot \frac{e^{\beta \cdot EV_{D_j}(t)}}{\sum_1^{N(j)} e^{\beta \cdot EV_{D_j}(t)}} + (1 - w_D) \cdot \frac{e^{\beta \cdot EV_{G_j}(t)}}{\sum_1^{N(j)} e^{\beta \cdot EV_{G_j}(t)}} \tag{3}$$

where $\beta = 3^c - 1 (0 \le c \le 5)$, and c is a log inverse temperature parameter that determines how consistently the option with the higher expected value is selected[35]. When $c = 0$, choices are random; as c increases, the option with the highest EV is selected more often. The EVs are the mathematically defined expectations for their respective distributions. In the multivariate Gaussian distribution, the EV for option j, $EV_{j,G}$, is the estimated mean, μ_j. In the Dirichlet distribution, $EV_{j,D}$ is the proportion of the estimated frequency, α_j, divided by the sum of all frequency estimates, $\frac{\alpha_j}{\sum_{i=1}^{k} \alpha_i}$.

We compared our dual-process model to five established RL models. The Delta and Decay models represent the best-performing examples from the two major classes of RL learning models with demonstrated effectiveness[9]. The Delta model[35–37], one of the most widely used mean-centered RL models, updates the EV of an option solely based on recency-weighted prediction errors and assumes no changes for unchosen options. In contrast, the Decay model[11,21], which excels in decision-making scenarios with unequal reward frequencies, adds the raw reward value to an option's EV but assumes that the EV decays for every time point the option is not chosen. We also compared our model to two uncertainty-sensitive models that account for reward variance. The risk-sensitive Delta model[24] follows the same updating rule as the Delta model but applies asymmetric learning rates for positive versus negative prediction errors. The mean-variance utility model[25], drawn from the economics literature, discounts an option's EV by its variance to incorporate risk sensitivity. Finally, we compared the dual-process model to a classic example from the sampler model class, the Adaptive Control of Thought—Rational (ACT-R) model[38,39]. We used a slightly modified version of the ACT-R model introduced by Erev and colleagues (2010) for applications in repeated RL tasks[8]. This ACT-R model probabilistically retrieves past experiences to guide future actions. It calculates the activation level for each previous experience of choosing an option and aggregates activated experiences to determine the EV using a probability-weighted sum. The computational mechanisms of these alternative models are detailed below.

Alternative models

The Delta rule[35–37] updates the EV by incorporating the prediction error between the EV from the last trial and the actual reward received in the current trial. EV_{t+1} for option j is defined as:

$$EV_{j,t+1} = EV_{j,t} + \alpha \cdot \left(r_t - EV_{j,t}\right) \cdot I_j \tag{4}$$

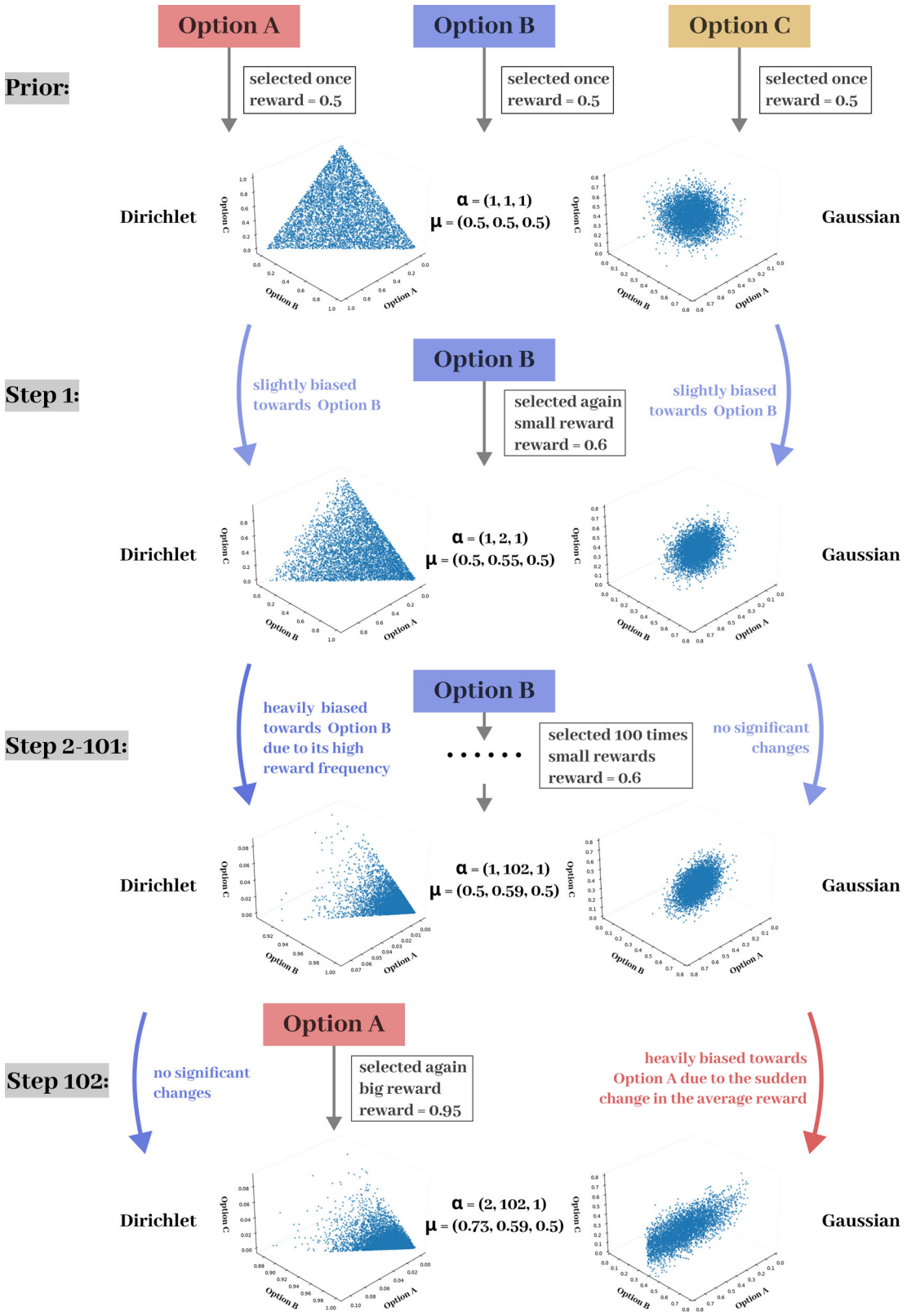
where I_j is a term recording option choice via a value of 1 if option j is chosen on trial t, and 0 otherwise; r_t is the reward value; and α is the learning parameter, $\alpha \in (0, 1)$, with higher α indicating greater emphasis on most recent samples. In this model, no memory of previous trials is retained, making it mean-centered. Consequently, when prediction errors are minimal, the EV will not increase significantly regardless of the number of rewards received.

The Decay model[11,21] updates the EV of option j on trial t as:

$$EV_{j,t+1} = EV_{j,t} \cdot (1 - A) + r_t \cdot I_j \tag{5}$$

where A is the decay parameter, $A \in (0, 1)$, with higher A indicating higher weights given to recent outcomes. In this model, the EV for each option will decay over time and only increase when a reward for that option is received. Thus, the more frequent the reward, the greater the EV, making the Decay model sensitive to reward frequencies.

The risk-sensitive Delta model[24] applies asymmetric learning rates for positive and negative prediction errors and then updates EV in the same



manner as the standard Delta model. Specifically,

$$EV_{j,t+1} = \begin{cases} EV_{j,t} + \alpha^{+} \cdot \delta(r_t) \cdot I_j & \text{if} \delta(r_t) > 0, \\ EV_{j,t} + \alpha^{-} \cdot \delta(r_t) \cdot I_j & \text{if} \delta(r_t) < 0, \end{cases} \quad (6)$$

where $\delta(r_t)$ represents the prediction error, $\delta(r_t) = r_t - EV_{j,t}$; α^{+} and α^{-} denote the learning rate for positive and negative prediction errors, respectively, $\alpha \in (0, 1)$. This model has been shown to effectively account for risk sensitivity in decision-making[24].

Fig. 2 | The updating mechanisms of the Dirichlet and Multivariate Gaussian distributions. This figure illustrates the basic mechanisms of our dual-process model, comparing how the Dirichlet and multivariate Gaussian distributions process reward information. Imagine we have three options: A, B, and C. As the prior, each option has been selected once and yielded a reward of 0.5. This results in equal probability mass being assigned to each option in both Dirichlet and Gaussian distributions, indicating no a priori preference. Next, if B is selected again and yields a reward of 0.6, both distributions become biased towards B because this increases both the tally of B yielding a reward and the average value of B so far. Now, if B is selected 100 times, with a reward of 0.6 each time, the tally of B yielding a reward above average becomes 102, causing the Dirichlet distribution to favor B almost exclusively. In contrast, the Gaussian distribution shows little change because it primarily processes the average value of B, which approximates 0.6 over time. In this scenario, the Dirichlet distribution predicts a close-to-1 probability of selecting B again, while the Gaussian distribution only gives a probability slightly higher than 0.5. Finally, imagine that on trial 102, A is selected and yields a reward of 0.95. This does not significantly affect the Dirichlet distribution, as the overwhelming advantage of B still exists. However, the Gaussian distribution changes considerably, now favoring A, as the average value for A increases significantly from 0.5 to 0.725, higher than 0.599 for B. At this point, the two distributions diverge in their predictions: the Dirichlet distribution continues to favor B very strongly, while the Gaussian distribution predicts a higher probability of choosing A over B due to its higher average reward value. Here, α denotes the vector for the number of successes and μ represents the vector for the mean values. For simplicity, neither the decay and learning rates nor the uncertainty are included here. The distribution plots represent the posterior Dirichlet and multivariate Gaussian distributions generated using the corresponding parameter sets. Blue dots indicate data points randomly sampled from the posterior distributions. For each posterior, 5,000 data points were randomly sampled, with the x-, y-, and z-axes representing the actual values of the sampled data points.

Another risk-sensitive model, the mean-variance utility model[25], discounts the EV of an option by its estimated variance. Specifically:

$$EV_j = \mu_j - \frac{\lambda \sigma_j^2}{2} \tag{7}$$

where μ and σ^2 represent the mean and variance of the outcomes for option j; and λ quantifies subjective risk aversion. A higher λ indicates greater risk aversion. In economics, this model typically excludes recency effects or the discounting of past outcomes. However, for its application in psychology, we update its mean and variance using a Delta rule:

$$\mu_{j,t+1} = \mu_{j,t} + \alpha \cdot \delta(r_t) \cdot I_j \tag{8}$$

$$\sigma^2_{j,t+1} = \sigma^2_{j,t} + \alpha \cdot \left(\delta(r_t)^2 - \sigma^2_{j,t}\right) \cdot I_j \tag{9}$$

where, again, $\delta(r_t)$ represents the prediction error, $\delta(r_t) = r_t - EV_{j,t}$; and α denotes the learning rate, $\alpha \in (0, 1)$.

Finally, the ACT-R model is a sampler model that represents a classic abstraction of the declarative memory system. As described in Erev et al. (2010), each trial is coded into an experience chunk that includes the participant's selection and the corresponding reward r. When option j is presented again, the agent considers all previous experiences selecting option j and recalls the experiences that exceed the activation level. The activation level of experience i is calculated as:

$$A_{j,i} = ln \sum_{k=1}^{n} t_{j,k}^{-\alpha} + \varepsilon(s) \tag{10}$$

where $t_{j,k}$ is the number of trials since the kth time option j was selected, α is the decay rate, and $\varepsilon(s)$ is a random value chosen from a logistic distribution with variance $\pi^2 s^2/3$, $s = \frac{1}{\sqrt{2}\cdot\beta}$. The random term implements a stochastic *SoftMax* retrieval process akin to Eq. 3 where the probability P_i of retrieving experience i is given by:

$$P_i = \frac{e^{A_{j,i}\cdot\beta}}{\sum e^{A_j\cdot\beta}} \tag{11}$$

where $\beta = 3^c - 1 (0 \le c \le 5)$. Then, the reward value of any experience that exceeds the activation level τ will be weighted by the corresponding P_i to calculate the weighted sum as the EV for option j. For $P_i > \tau$, the EV is calculated as:

$$EV_j = \sum_{i \in \{i | A_i > \tau\}} P_i \cdot r_i \tag{12}$$

where r_i is the actual reward value for experience i. This method allows the ACT-R model to stochastically integrate past experiences and predict the probability of selecting option j based on the weighted contributions of all relevant experiences.

Model Simulation

To evaluate how each of the six models would adapt to various RL scenarios, we randomly sampled 5000 reward schedules while maintaining the same basic task structure: (1) C > A > B > D, with at least a 0.01 mean difference; (2) $C, A \in (0.5, 1)$ and $B, D \in (0, 0.5)$; 3) AB trials presented twice as often as CD trials during the training phase; (3) reward standard deviation $\in (0.11, 0.48)$. For each reward schedule, we ran 1000 simulations to predict preferences in CA trials. Across a wide range of reward ratios and variances, we generated two sets of model surfaces: one showing the predicted proportion of C choices in CA trials and another showing the proportion of simulated agents displaying a disproportionate preference for the more frequently rewarded option A. The reward ratio between two options is calculated as $\frac{\mu_1}{\mu_1+\mu_2}$, and choosing C less frequently than this ratio indicates a disproportionate preference for A. Previous research[40] demonstrated that simulating the entire surface of computational models can reveal underlying model properties and task dynamics that might be missed when focusing solely on specific reward schedules.

To avoid simulation result variations caused by differing schedules, the same 5000 generated reward schedules were applied to all models. During these simulations, parameters were randomly drawn from uniform distributions. The learning rate α and decay parameter A values were bounded between 0 and 1, c between 0 and 5, τ between -2 and 0, and all weight parameters, including λ, were constrained between 0 and 1[11,35]. Randomized trial sequences in both training and transfer phases were generated and shuffled for simulation. For traditional model simulations presented in the supplementary materials, we followed the same procedure but increased the sample size, simulating 10,000 virtual agents for each condition.

In the post-hoc simulation, we randomly sampled 10,000 parameter sets with replacement from the individualized best-fitting parameters obtained during model fitting (akin to bootstrapping). These sampled best-fitting parameters were then used to simulate the model with randomized trial sequences, allowing us to estimate the model's ability to replicate empirical behavioral patterns. For each model, we calculated the percentage of choosing the optimal option in each pair (e.g., A in AB trials, C in CA trials) as predicted by the simulations. The root mean squared error (RMSE) was then computed between the actual and predicted percentage of choosing the optimal option across six trial types to estimate model performance.

Model fitting and evaluation

We used the maximum likelihood (ML) approach for model fitting. The negative log likelihood of the parameter set θ, given observed data y and model M, $L(\hat{\theta}|y, M)$, is minimized using the *minimize* function in the *SciPy* library in Python. To avoid local minima, we ran the optimization 200 times

for each participant with randomly selected starting points for each parameter. Data from both training and transfer phases were used during model fitting.

We computed the Akaike Information Criteria (AIC)[41] and Bayesian Information Criteria (BIC)[42] for each individual participant for each model, and used these indices to calculate AIC and BIC weights for model comparison[43]. AIC is calculated as:

$$AIC = -2\ln L\left(\widehat{\theta}, |, y, M\right) + 2K \tag{13}$$

where K is the number of free parameters in the model. BIC is calculated as:

$$BIC = -2\ln L\left(\widehat{\theta}, |, y, M\right) + K\ln(N) \tag{14}$$

where N is the number of observations. In our study, N equals the 250 trials for each participant. The AIC and BIC weights are calculated as:

$$w_i(\mathrm{AIC/BIC}) = \frac{\exp -\frac{1}{2}\Delta_i(\mathrm{AIC/BIC})}{\sum_{k=1}^{K}\exp -\frac{1}{2}\Delta_k(\mathrm{AIC/BIC})} \tag{15}$$

where $\Delta_i(\mathrm{AIC/BIC}) = \mathrm{AIC}_i/\mathrm{BIC}_i - \min(\mathrm{AIC/BIC})$. Higher AIC or BIC values indicate worse model fit, whereas higher AIC- and BIC-weights indicate greater support for the model relative to all comparison models. We also calculated the Bayes Factor (BF_{10}) as $BF_{10,\ Model1} = \exp\left(\frac{BICmodel2 - BICmodel1}{2}\right)$[44]. A BF_{10} higher than 3 is generally considered significant, representing a moderate sized effect[44].

In addition, we applied Variational Bayesian Model Selection (VBMS)[45] criteria, which are designed for group-level model comparisons. VBMS treats the model as a random variable and estimates the parameters of a Dirichlet distribution, which are then used to define a multinomial distribution that provides the probability of each model generating the data for a randomly selected subject. Specifically, the posterior Dirichlet parameters, α, represent the estimated frequency of each model being the one that generates the data for a given subject. The posterior multinomial parameter, r_k, describes the probability that data from a randomly chosen subject is generated by a specific model k. Finally, the exceedance probability, φ_k, quantifies the likelihood that a particular model k is more likely than any other model in the set to generate group-level data. We used BIC to approximate the log evidence (see Supplementary Table 3 for results using AIC as the log evidence) and all three metrics were calculated for model comparisons.

Model-recovery and parameter-recovery

To assess the significance of the findings obtained in the model fitting analyses, we conducted model recovery and parameter recovery analyses[46]. These analyses assessed whether the data generated by the dual-process model could be accurately recovered by itself and whether the fitting process could retrieve the underlying parameter set that generated the data. Briefly, we simulated datasets using all six main models: the dual-process model, Delta, risk-sensitive Delta, mean-variance utility, Decay, and ACT-R. Each simulated dataset was then fit by all models to assess how well the data generated by Model A can be best fit (i.e., recovered) by Model A compared to other models. This process was repeated 100 times for each condition, with 30 random starting parameter sets used during the fitting of the simulated data. Since we were testing a new model and the "true" parameter range was largely unknown, we did not manually set any additional boundaries on the parameters. Therefore, all simulations for all models were conducted using parameters drawn from their entire possible range.

The results were presented in two matrices. The confusion matrix reports the probability that Model A fits the data best given the data was generated by Model A, P (fit model | simulated model), while the inversion matrix reports the probability that Model A generated the data given it was best fit by Model A, P (simulated model | fit model). For both matrices, a higher score indicates better recovery performance.

This simulation process also provided insights into parameter recovery. For this, the underlying parameter sets used during simulation were recorded and correlated with the best-fitting parameter sets recovered by the same model. Ideally, the data generated by a Model A with given parameters can be fit by Model A to recover those parameters, leading to a high correlation between the generating parameter set and the recovered parameter set[46]. A higher correlation indicates better parameter recovery.

Results

A priori model simulations

As shown in Fig. 3a, the dual-process model was the only model sensitive to both reward ratio and variance: it predicted fewer C choices in CA trials with both higher variance and lower reward ratio. In contrast, the Delta, risk-sensitive Delta, mean-variance utility, Decay and ACT-R models showed a sharp increase in the proportion of C choices as reward ratio increased, whereas their surfaces remained nearly flat across the variance axis. This indicates that these models primarily focused on reward value differences while ignoring the impact of uncertainty. Similarly in Fig. 3b, the dual-process model predicted an increasing proportion of simulated agents displaying frequency effects with both higher reward variance and lower reward ratio. The Decay model predicted frequent occurrences of frequency effects overall, even with high reward ratios, but showed little sensitivity to changes in reward variance. The other four models predicted minimal changes as reward variance varied. In our paradigm, the two uncertainty-sensitive models failed to adapt to different levels of reward variance because they discounted the EV of all options equally when all options had the same level of reward variance. This uniform discounting prevented relative EV differences from changing with altered reward variance, as these models computationally overlooked the role of reward frequency.

Statistically, this pattern was confirmed through logistic regression models. Logistic regression predicting the proportion of C choices by reward ratio and variance (*Proportion of C Choices ~ Reward Variance × Reward Ratio*) showed that, across all six models, the proportion of C choices in CA trials generally had a positive relation with the reward ratio (dual-process: $b = 13.581 \pm 2.561$, $z = 5.303$, $p < 0.001$, 95% CI = [8.599, 18.641]; Delta: $b = 18.537 \pm 2.914$, $z = 6.361$, $p < 0.001$, 95% CI = [12.882, 24.309]; risk-sensitive Delta: $b = 18.494 \pm 2.913$, $z = 6.348$, $p < 0.001$, 95% CI = [12.841, 24.264]; mean-variance utility: $b = 18.662 \pm 2.908$, $z = 6.418$, $p < .001$, 95% CI = [13.020, 24.421]; Decay: $b = 7.494 \pm 2.360$, $z = 3.175$, $p = .001$, 95% CI = [2.875, 12.130]; ACT-R: $b = 17.774 \pm 3.149$, $z = 5.644$, $p < 0.001$, 95% CI = [11.671, 24.020]), indicating more C choices as reward ratio increased. However, only in the dual-process model did the regression model find a significant interaction effect, where increased reward variance attenuated the positive association between reward ratio and C choices ($b = -15.826 \pm 7.907$, $z = -2.002$, $p = 0.045$, 95% CI = [−31.373, −0.368]). For the other five models, the interaction effects were non-significant (Delta: $b = -15.144 \pm 8.960$, $z = -1.690$, $p = 0.091$, 95% CI = [−32.742, 2.391]; risk-sensitive Delta: $b = -15.093 \pm 8.956$, $z = -1.685$, $p = 0.092$, 95% CI = [−32.682, 2.434]; mean-variance utility: $b = -16.009 \pm 8.933$, $z = -1.792$, $p = 0.073$, 95% CI = [−33.556, 1.473]; Decay: $b = 0.145 \pm 7.472$, $z = 0.019$, $p = 0.985$, 95% CI = [−14.508, 14.792]; ACT-R: $b = -7.692 \pm 9.701$, $z = -0.793$, $p = 0.428$, 95% CI = [−26.740, 11.300]).

A second set of logistic regression models predicting the proportion of simulated agents showing frequency effects by reward ratio and variance (*Proportion of Frequency Effects ~ Reward Variance × Reward Ratio*) revealed similar results. While all models, except for the Decay model, predicted a lower prevalence of frequency effects as reward ratio increased (dual-process: $b = -13.456 \pm 2.505$, $z = -5.371$, $p < 0.001$, 95% CI = [−18.405, −8.581]; Delta: $b = -16.633 \pm 2.809$, $z = -5.920$, $p < 0.001$, 95% CI = [−22.194, −11.178]; risk-sensitive Delta: $b = -16.566 \pm 2.810$, $z = -5.895$, $p < 0.001$, 95% CI = [−22.129, −11.110]; mean-variance utility: $b = -16.649 \pm 2.801$, $z = -5.944$, $p < 0.001$, 95% CI = [−22.194, −11.210]; Decay: $b = -3.182 \pm 2.417$, $z = -1.316$, $p = 0.188$, 95% CI = [−7.915, 1.563];

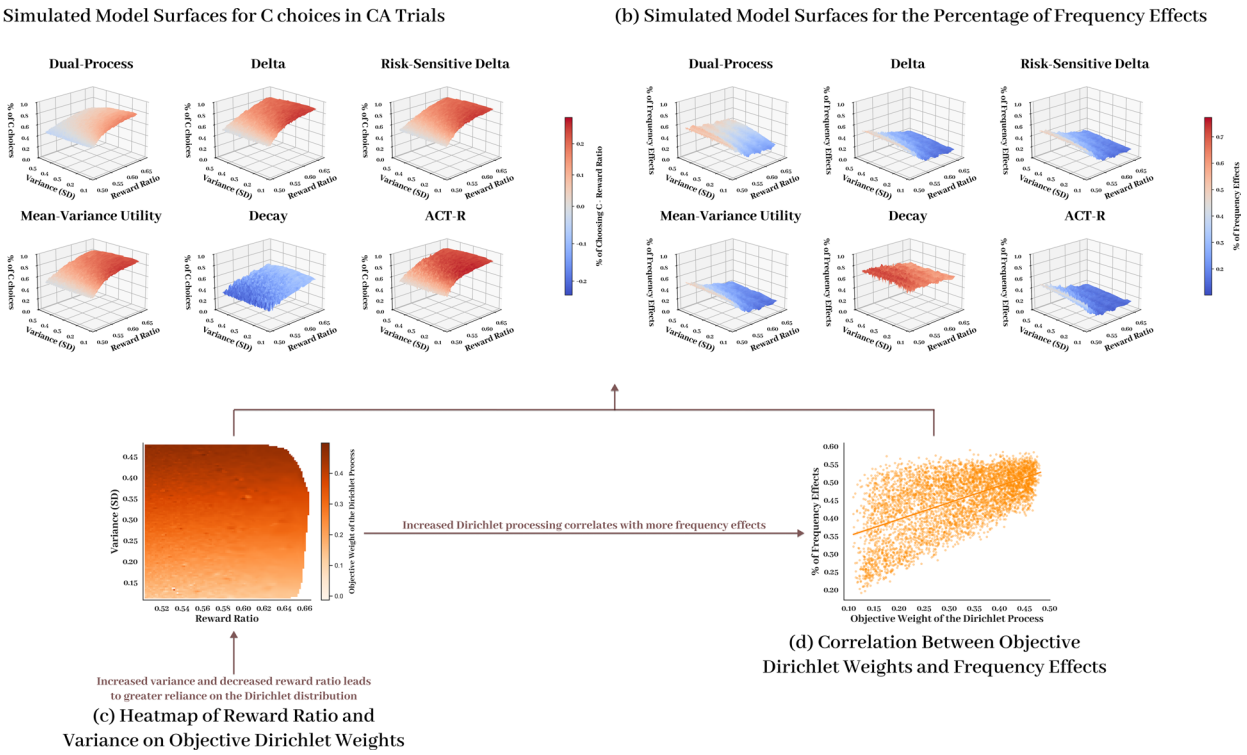


Fig. 3 | Model simulation results. a Simulated model surfaces showing the proportion of C choices in CA trials reveal that the dual-process model uniquely demonstrates sensitivity to both increased variance and lower reward ratios. In contrast, the other five models exhibit sharp declines in frequency effects as reward ratio increases but remain largely insensitive to variance. Color represents the difference between the reward ratio and the C choice rate, with blue indicating frequency effects (i.e., a preference for the more frequently rewarded option A) and red indicating a stronger preference for C than predicted by the reward ratio. **b** Model surfaces for the proportion of simulated agents showing frequency effects further highlight that the dual-process model uniquely predicts a higher prevalence of frequency effects with both increased variance and lower reward ratios. The Delta, risk-sensitive Delta, mean-variance utility, and ACT-R models again show sharp declines in frequency effects as reward ratio increases while remaining unresponsive to changes in variance. The Decay model, while consistently predicting strong frequency effects, also displays limited sensitivity to variance (see also Supplementary Figs. 1, 2 for traditional simulations limited to the three experimental conditions). The color bar is normalized to represent the percentage of simulated agents displaying frequency effects. **c** The heatmap shows that higher reward variance and lower reward ratios result in a stronger reliance on the Dirichlet distribution. **d** The scatterplot shows that increased objective Dirichlet weights are positively correlated with greater preference for A in CA trials, supporting the hypothesis that greater uncertainty enhances reliance on frequency-driven processes, which is directly related to more prevalent frequency effects. In these figures, "variance" conceptually refers to reward variance rather than statistical variance. As indicated in the brackets, the values represent the standard deviation of the normal distributions from which the rewards are drawn.

ACT-R: $b = -15.542 \pm 3.108$, $z = -5.000$, $p < 0.001$, 95% CI = [−21.705, −9.518]), only the dual-process model predicted that this negative relationship would be mitigated by increased reward variance (dual-process: $b = 23.913 \pm 7.730$, $z = 3.094$, p =0.002, 95% CI = [8.816, 39.126]; Delta: $b = 16.278 \pm 8.612$, $z = 1.890$, $p = 0.059$, 95% CI = [−0.570, 33.200]; risk-sensitive Delta: $b = 16.098 \pm 8.612$, $z = 1.869$, $p = 0.062$, 95% CI = [−0.750, 33.644]; mean-variance utility: $b = 16.774 \pm 8.585$, $z = 1.954$, $p = 0.051$, 95% CI = [−0.020, 33.644]; Decay: $b = -0.535 \pm 7.663$, $z = -0.070$, $p = 0.944$, 95% CI = [−15.557, 14.495]; ACT-R: $b = 7.902 \pm 9.482$, $z = 0.833$, $p = .405$, 95% CI = [−10.648, 26.531]).

We closely examined the underlying cause of dual-process model's sensitivity to reward variance. As predicted, higher variance increases the statistical dispersion in the Gaussian process, which in turn reduces its objective weight. In contrast, the Dirichlet process remains relatively unaffected, as the frequency with which simulated agents receive above-average rewards from options A and C does not vary significantly across different levels of variance. Assuming a uniformly distributed personal preference for the two processes, the decreased objective weight of the Gaussian process results in a higher overall weight for the Dirichlet process, which, in turn, favors A and leads to fewer selections of option C in CA trials.

Behavioral results

To validate our model predictions, 293 participants completed the task (LV: 93; MV: 100; HV: 100). Figure 4a shows the proportion of optimal choices for each trial type during both training and transfer phases. Logistic regression analyses (*Proportion of Optimal Option ~ Condition + Trial Type*) indicated that participants were generally less likely to select the optimal option in the MV condition compared to the LV condition ($b = -0.273 \pm 0.023$, $z = -11.892$, $p < 0.001$, 95% CI = [−0.228, −0.318]), and in the HV condition compared to the MV condition ($b = -0.518 \pm 0.021$, $z = -25.020$, $p < 0.001$, 95% CI = [−0.559, −0.478]) across all trial types, indicating poorer learning with increased uncertainty. During training, we observed slower increases in optimal choices (i.e., A and C) with higher variance (MV-LV: $F_{5,955} = 2.34$, $p = .040$, $\eta_p^2 = 0.012$; HV-MV: $F_{5,990} = 15.76$, $p < 0.001$, $\eta_p^2 = 0.074$; Supplementary Fig. 3). This could reflect a classic exploitation-exploration tradeoff in time-constrained scenarios, where participants must balance efforts to take advantage of known rewards with exploring the unknown environment[16,47]. Under low uncertainty, participants quickly identified and committed to better options, whereas under high uncertainty, they were more motivated to explore, as no clear target for exploitation emerged[47,48].

For the critical CA trials, shown in Fig. 4b, we conducted one-sample *t* tests to compare the proportion of C choices against a chance level of 0.5 and the objective reward ratio between the two alternatives. In CA trials, the reward ratio is 0.75/(0.75+0.65)=0.536 in favor of option C. Intriguingly, participants showed a significant preference for the theoretically better option C in the LV condition (chance-level: $t(92) = 3.051$, $p = 0.003$, $d = 0.316$; reward-ratio: $t(92) = 2.107$, $p = 0.038$, $d = 0.219$; 95% CI = [0.540,

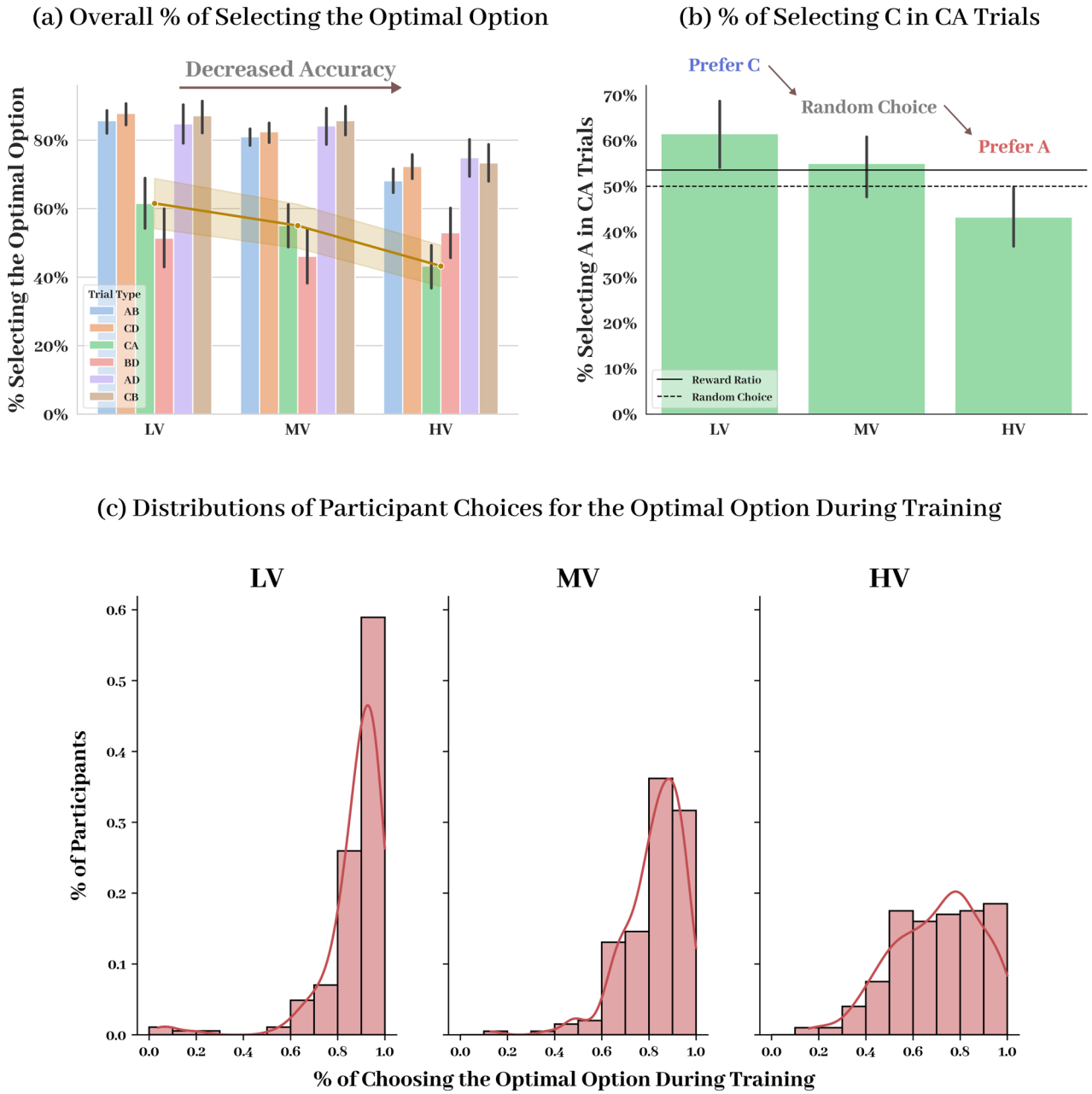


Fig. 4 | Behavioral results. a Participants in the LV condition showed the highest accuracy in selecting the optimal option during training, with performance decreasing as uncertainty increased in MV and HV conditions. This trend suggests that greater uncertainty led to poorer learning, likely due to an exploration-exploitation tradeoff where participants were more inclined to explore in high-uncertainty scenarios rather than commit to a known optimal option. Here, the optimal option refers to the option whose distribution, from which the outcomes are drawn, has a higher mean and thus provides better long-term payoffs with infinite draws. Accordingly, the ranking is C > A > B > D. **b** During CA trials, participants showed a preference for the better option C in the LV condition, no clear preference in the MV condition, and a preference for the less rewarding but more frequently rewarded option A in the HV condition. These results align with the predictions of the dual-process model, indicating that unequal reward frequencies have a stronger influence when reward uncertainty is higher. Dashed line refers to the reward ratio (0.536) while solid line refers to the random chance (0.5) choice rate. **c** Histograms across LV, MV, and HV conditions show that participants' distribution of choices shifted with increasing uncertainty. Under low variance, participants favored the optimal option more consistently, while in the HV condition, choices were more dispersed, reflecting greater exploration and less commitment to the optimal choice. Error bar represents 95% confidence interval. $N = 93$ participants in the LV condition, 100 participants in the MV condition, and 100 participants in the HV condition.

0.691]), no statistically significant preference for either option in the MV condition (chance-level: $t(99) = 1.513$, $p = 0.134$, $d = 0.151$; reward-ratio: $t(99) = 0.441$, $p = 0.660$, $d = 0.044$; 95% CI = [0.484, 0.617]), and a significant preference for the less rewarding but more frequently rewarded option A in the HV condition (chance-level: $t(99) = -2.133$, $p = 0.035$, $d = 0.213$; reward-ratio: $t(99) = -3.260$, $p = 0.002$, $d = 0.326$; 95% CI = [0.370, 0.495]). These results map nicely onto the predictions of our dual-process model by indicating that, as hypothesized, the impact of unequal reward frequency on decision-making escalates with increased value uncertainty. When the underlying value is harder to gauge, individuals are more likely to rely on their intuitive sense of how frequently an option has yielded an above-average reward.

Model fitting results

Table 2 presents the model fitting results. Across all reward variance levels, the dual-process model consistently demonstrated a substantial advantage

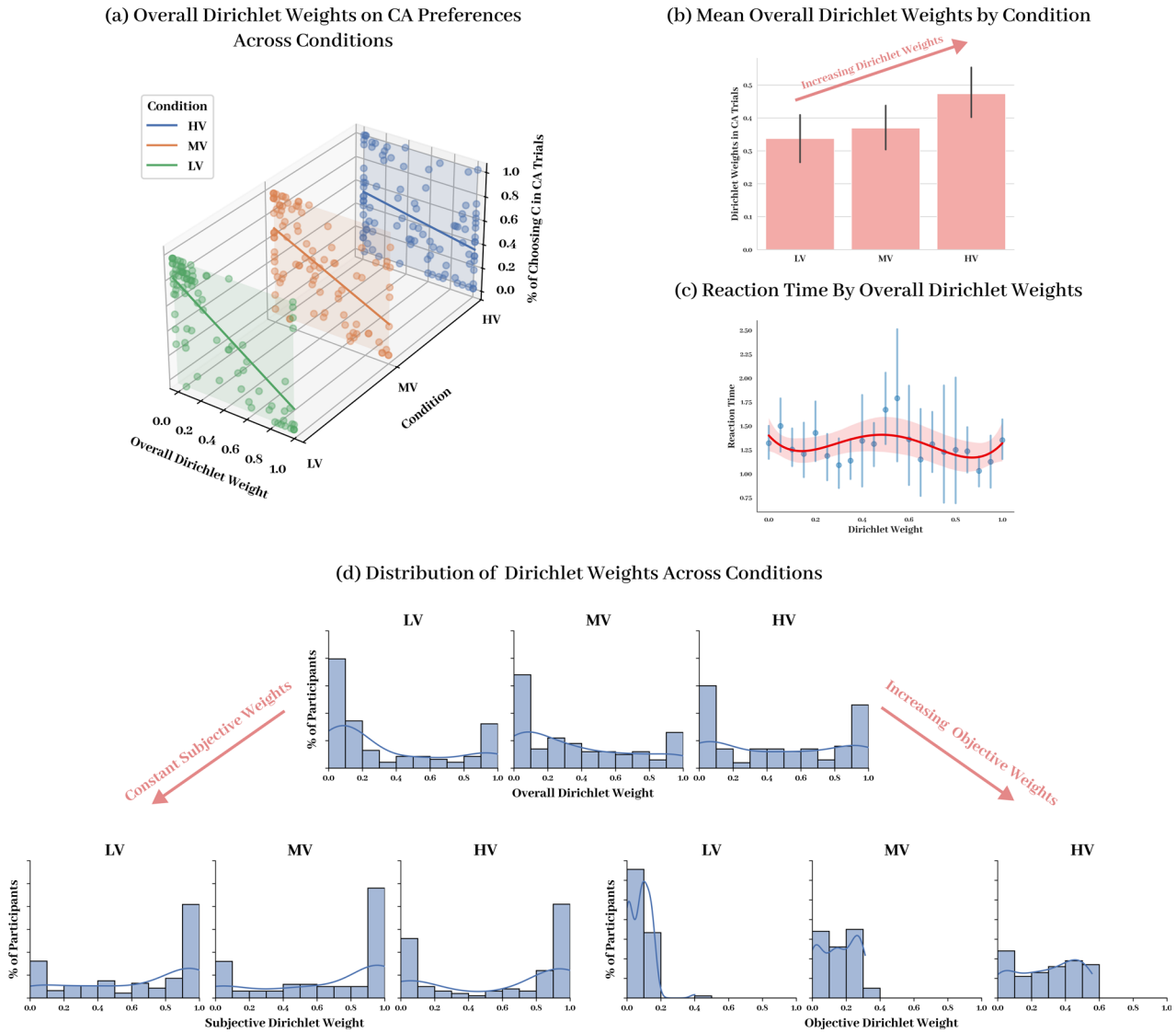


Fig. 5 | Behavioral results explained by model-inferred Dirichlet weights. a A 3-dimensional scatterplot reveals that higher Dirichlet weights were consistently associated with a greater likelihood of choosing option A in CA trials, empirically validating the predictions of the dual-process model. **b** Bar plots show that overall Dirichlet weights significantly increased from LV to HV conditions, supporting the hypothesis that greater variance leads to increased reliance on frequency-based processing. **c** Reaction times, modeled as a quartic regression, indicate that participants experienced the greatest cognitive load when Gaussian and Dirichlet processes had nearly equal weights and both required careful evaluation. **d** Histograms show that subjective Dirichlet weights remained nearly constant across conditions, while objective Dirichlet weights increased with variance, highlighting environmental uncertainty as the key factor driving the shift toward Dirichlet frequency-based processing. This pattern is hypothesized to be caused by the combined influence of unequal reward frequency and increasing reward variance, while we found that increased reward variance alone was sufficient to produce a similar pattern of Dirichlet weight changes (see Supplementary Fig. 8). Error bar represents 95% confidence interval. $N = 93$ participants in the LV condition, 100 participants in the MV condition, and 100 participants in the HV condition.

over all other models, with the only exception of the risk-sensitive Delta model in the MV condition, where performance was comparable. In the HV condition, it had a mean AIC advantage of 14.06 and a mean BIC advantage of 12.05 over the other five well-established RL models, corresponding to a BF_{10} of 65.75 over Delta, 8.89 over risk-sensitive Delta, 111.01 over mean-variance utility, 99.40 over Decay, and 9.918×10^6 over ACT-R. This advantage was equally large in the MV (AIC-advantage: 14.71; BIC-advantage: 12.70) and LV (AIC-advantage: 18.18, BIC-advantage: 16.17) conditions, which cannot be attributed to either of the individual processes alone (Supplementary Table 1). Notably, in addition to the six previously described models, we also fit a purely objective model where the distributional weights were entirely determined by the objective weight. We found that even the purely objective version of the dual-process model fits better than many current RL models, suggesting that simply accounting for the weighting of multiple processes by their respective entropy—without considering individual differences—was already sufficient to surpass models that assume a single decision-making strategy (Supplementary Table 1).

Post-hoc simulations using the individualized best-fitting parameters also showed that the dual-process model has the lowest RMSE among all models across conditions (dual-process: .029; Delta: 0.040; risk-sensitive Delta: 0.044; mean-variance utility: 0.044; Decay: 0.095; ACT-R: 0.052; see Supplementary Table 4 for details). The dual-process model replicated participants' choice patterns in CA trials most successfully, whereas the Decay model predicted consistent frequency effects irrespective of reward variance, and the remaining four models failed to recover any frequency effects during post-hoc simulations across the three levels of reward variance (Supplementary Fig. 4).

Recovery results showed that the dual-process model successfully recovered 67% of data generated by itself in the LV condition, 75% in the MV condition, and 81% in the HV condition (Supplementary Fig. 5). All

https://doi.org/10.1038/s44271-025-00249-y

Table 2 | Model fitting results

	Best c	Best $\alpha/\alpha^{+}/A$	Best $w_1/\tau/\lambda/\alpha^{-}$	AIC_{avg}	BIC_{avg}	AIC_{weight}	BIC_{weight}	BF_{10}	VB α	VB r_k	VB φ_k
LV											
Delta	2.556	0.320		186.520	193.563	<0.001	0.001	732.959	12.629	0.128	<0.001
Risk-Sensitive Delta	2.742	0.316	0.310	175.397	185.962	0.057	0.057	16.389	15.830	0.160	<0.001
Mean-Variance Utility	2.324	0.241	12.424	182.421	192.985	0.002	0.002	549.121	11.031	0.111	<0.001
Decay	0.935	0.180		197.625	204.668	<0.001	<0.001	1.891×10^5	11.636	0.118	<0.001
ACT-R	1.952	0.483	–0.844	191.576	202.140	<0.001	<0.001	5.342×10^4	2.965	0.030	<0.001
Dual-Process	***2.356***	***0.301***	***0.688***	***169.804***	***180.368***	***0.941***	***0.940***	**–**	***44.909***	***0.454***	***0.999***
MV											
Delta	2.379	0.322		212.609	219.652	0.002	0.012	36.492	18.649	0.176	0.027
Risk-Sensitive Delta	2.339	0.399	0.314	201.217	211.782	0.594	0.589	0.713	28.780	0.272	0.581
Mean-Variance Utility	2.227	0.306	7.235	210.062	220.627	0.007	0.007	59.394	9.676	0.091	<0.001
Decay	0.949	0.204		224.370	231.413	<0.001	<0.001	1.306×10^4	19.352	0.183	0.036
ACT-R	1.641	0.491	-0.977	226.832	237.396	<0.001	<0.001	2.601×10^5	2.997	0.028	<0.001
Dual-Process	***2.302***	***0.321***	***0.627***	***201.894***	***212.458***	***0.413***	***0.409***	**–**	***26.546***	***0.250***	***0.355***
HV											
Delta	1.784	0.381		264.439	271.482	0.002	0.013	65.748	22.846	0.216	0.085
Risk-Sensitive Delta	1.774	0.357	0.385	256.916	267.48	0.010	0.098	8.890	13.557	0.128	<0.001
Mean-Variance Utility	1.838	0.398	11.305	261.965	272.53	0.008	0.008	111.011	10.095	0.095	<0.001
Decay	0.876	0.264		265.266	272.309	0.002	0.009	99.404	24.899	0.235	0.162
ACT-R	0.971	0.422	–1.112	284.766	295.33	<0.001	<0.001	9.918×10^6	3.011	0.028	<0.001
Dual-Process	***2.019***	***0.385***	***0.602***	***252.546***	***263.110***	***0.888***	***0.872***	**–**	***31.593***	***0.298***	***0.752***

Model Fitting Results. This table summarizes the model fitting results (n = 293 participants). The parameter c is the inverse log temperature parameter. A higher c means the participant is more likely to stick with the option that has a theoretically higher EV. We made a slight modification to the original implementation of the ACT-R model[8] by applying the same version of the *SoftMax* rule used in the other five models. This modification did not significantly affect model fit. The α or A parameter represents the level of recency effects, indicating how strongly people rely on recent outcomes to make their next decision. A higher α or A signifies more rapid decay of past experiences and greater reliance on recent samples. The third parameter in the risk-sensitive Delta model, α^{-}, represents the asymmetric learning rate for negative prediction errors. In the mean-variance utility model, the third parameter, λ, represents subjective utility curve, with a higher λ indicating higher risk-aversion. In the ACT-R model, τ represents the memory retrieval threshold, where only past experiences with an activation level above τ can be retrieved and considered in the calculation of EV. AIC, BIC, AIC-weight and BIC-weight represent indices for model fit. Lower AIC and BIC values indicate better model fit, while higher AIC- and BIC-weights suggest a stronger relative advantage for the corresponding model. AIC- and BIC-weights sum to one. BF_{10} presents the BF difference between the corresponding model and our dual-process model. Typically, an AIC or BIC difference of 0-2 means little support for the better model, 4-7 indicates moderate support, and a difference of 10 or more signifies substantial support. A BF_{10} greater than 3 is considered significant (see Supplementary Table 2 for full BF tables). Finally, the last three columns present results for VBMS results with BIC being used as the log evidence.

three model parameters demonstrated significant correlations between the best-fitting parameters and the generating parameters, indicating satisfactory model and parameter recovery performance (c: Pearson $r = .905$, $p < 0.001$, 95% CI = [0.882, 0.924]; α: Pearson $r = 0.836$, $p < 0.001$, 95% CI = [0.798, 0.867]; w_1: Pearson $r = 0.209$, $p < 0.001$, 95% CI = [0.315, 0.098]; Supplementary Fig. 6).

Post hoc analysis

Now that we know the dual-process model outperforms many existing RL models across different reward variance levels, what does this tell us about the trial-by-trial decision-making process? To explore this, we closely examined participants' best-fitting parameters, weight distributions between C and A, and their ultimate choices in CA trials.

General linear model (GLM) analyses (*Model Parameter ~ Condition*) revealed that, in the dual-process model, the inverse log temperature parameter (c) significantly decreased from LV to HV ($b = –0.337 \pm 0.087$, $t = –3.881$, $p < 0.001$, 95% CI = [–0.507, –0.167]) and from MV to HV ($b = –0.282 \pm 0.085$, $t = –3.313$, $p = 0.001$, 95% CI = [–0.449, –0.115]), but not from LV to MV ($b = –0.055 \pm 0.087$, $t = –0.629$, $p = 0.530$, 95% CI = [–0.225, 0.116]). The decreased c with increased reward variance was consistent across models (*Model Parameter ~ Condition + Model*; LV-MV: $b = –0.176 \pm 0.063$, $t = –2.793$, $p = 0.005$, 95% CI = [–0.300, –0.053]; MV-HV: $b = –0.448 \pm 0.062$, $t = –7.222$, $p < 0.001$, 95% CI = [–0.570, –0.326]), indicating a general increase in random choice behavior from LV to HV as captured by all RL models. The learning/decay rate (α) did not significantly vary across conditions in the dual-process model (LV-MV: $b = 0.020 \pm 0.054$, $t = 0.364$, $p = 0.716$, 95% CI = [–0.086, 0.125]; MV-HV: $b = 0.064 \pm 0.053$, $t = 1.212$, $p = 0.227$, 95% CI = [–0.039, 0.167]; LV-HV: $b = 0.083 \pm 0.054$, $t = 1.553$, $p = 0.121$, 95% CI = [–0.022, 0.189]). Similarly, the subjective Dirichlet weight parameter (w_1) showed no significant differences between conditions (LV-MV: $b = –0.061 \pm 0.056$, $t = –1.093$, $p = 0.275$, 95% CI = [–0.171, 0.048]; MV-HV: $b = –0.024 \pm 0.055$, $t = –0.444$, $p = .657$, 95% CI = [–0.132, 0.083]; LV-HV: $b = –0.085 \pm 0.056$, $t = –1.529$, $p = 0.127$, 95% CI = [–0.195, 0.024]).

To explore the relationship between the involvement of the Dirichlet process and the observed preference shift, we conducted a multilevel mixed-effects logistic regression to predict the probability of choosing the optimal option C based on the overall weight of the Dirichlet process in CA trials, *Choosing C ~ Dirichlet weight + Condition + (1|Participant)*. After adjusting for condition, we found that higher Dirichlet weights were closely associated with a lower likelihood of selecting C ($b = –3.071 \pm 0.358$, $z = –8.569$, $p < 0.001$, 95% CI = [–3.815, –2.374]), suggesting a direct link between Dirichlet-based processing and frequency effects. Importantly, this effect could not be mediated by variations in any other parameters or misestimations in the Gaussian value-based process (Supplementary Table 5 & Supplementary Fig. 7). Further mixed-effects analyses, *Dirichlet Weight ~ Condition + (1|Participant)*, confirmed that the overall Dirichlet weights during CA trials significantly increased from LV to HV ($b = 1.242 \pm 0.546$, $z = 2.274$, $p = 0.023$, 95% CI = [0.169, 2.321]), and numerically but not significantly from LV to MV ($b = 0.210 \pm 0.544$, $z = 0.386$, $p = 0.670$, 95% CI = [–0.864, 1.281]), and MV to HV ($b = 1.032 \pm 0.536$, $z = 1.926$, $p = 0.054$, 95% CI = [–0.020, 2.093]). These combined effects align with our simulation results, demonstrating that increased variance leads to higher Dirichlet weights, which, in turn, encourage the selection of option A in CA trials due to its more frequent rewards.

Next, we decoupled the overall Dirichlet weights in CA trials to isolate the effects of objective versus subjective weights. As shown in Fig. 5, promisingly, two-sample Kolmogorov-Smirnov tests indicated that the distribution of subjective Dirichlet weights did not significantly vary across conditions (LV-MV: $D = 0.165$, $p = 0.124$; MV-HV: $D = 0.120$, $p = 0.468$; LV-HV: $D = 0.148$, $p = 0.212$), with nearly identical bimodal distributions. In contrast, the objective Dirichlet weights (*Objective Dirichlet Weights ~ Condition + Trial Type*) significantly increased from LV to MV ($b = 0.525 \pm 0.027$, $z = 19.736$, $p < 0.001$, 95% CI = [0.473, 0.577]), and MV to HV ($b = 0.732 \pm 0.022$, $z = 34.011$, $p = 0.022$, 95% CI = [0.690, 0.775]). This increase was even more pronounced in CA trials (LV-MV: $b = 0.721 \pm 0.095$, $z = 7.587$, $p < 0.001$, 95% CI = [0.536, 0.909]; MV-HV: $b = 0.836 \pm 0.071$, $z = 11.732$, $p < 0.001$, 95% CI = [0.697, 0.976]). By examining the distribution of objective weights, we found that very few participants ever considered the Dirichlet process when the reward variance was low. However, as uncertainty increased, reliance on reward frequency grew proportionally. This indicates that as variance rose, participants became less able to rely on a single decision-making strategy. The increased uncertainty prompted them to consult multiple estimation methods, such as using reward frequency as a proxy for value.

Lastly, we examined reaction times between Dirichlet- and Gaussian-oriented decisions during CA trials to infer the mental effort involved in utilizing different decision-making strategies. A GLM analysis (*Reaction Time ~ Dirichlet Weight + Condition*) indicated that participants generally took longer to decide in the HV condition than in the LV condition ($b = 0.155 \pm 0.053$, $t = 2.921$, $p = 0.004$, 95% CI = [0.051, 0.258]). A quartic version of the model revealed that Gaussian-oriented decisions required slightly longer processing times than Dirichlet-oriented decisions ($b = –3.058 \pm 1.812$, $t = –1.688$, $p = 0.092$, 95% CI = [–6.610, 0.494]), whereas participants hesitated the most when the decision involved both processes with nearly equal weights ($b = 5.113 \pm 1.812$, $t = 2.822$, $p = 0.005$, 95% CI = [1.562, 8.665]). This suggests that value-based decisions, which are likely to involve more rigorous calculations and EV estimations, might be more computationally demanding than frequency-based decisions, but the cognitive load is at its highest when multiple processes must be carefully considered before making the decision. We also observed slight rebounds in reaction time when the weight approached either extreme (i.e., Dirichlet weight close to 0 or 1)—in other words, when a decision was modeled as being almost entirely reliant on one decision-making system. This pattern suggests that individuals might experience tension when relying solely on one system, prompting a brief re-evaluation of the decision. However, this explanation is highly speculative, and further empirical support is needed.

Discussion

While it has been increasingly acknowledged that using a fixed learning rule to generate singular EVs provides a limited view on how people adjust their expectations[49–51], developing alternative approaches to quantitatively theorize behavioral decision-making remains challenging. For this purpose, our parallel distributional model provides a novel and parsimonious framework for understanding decision-making, particularly in complex, multifaceted scenarios. Deeply rooted in recent accounts of the "Bayesian brain" hypothesis[13,14,18,19], distributional representations of expectations have achieved notable success in explaining a wide range of human behaviors—many of which, just like frequency effects, appear irrational[52,53]. We extended this approach by demonstrating that distributional models can accommodate multiple dissociable estimation processes, thereby becoming highly generalizable across a variety of decision-making contexts.

In our decision-making task with three levels of reward variance, we found that participants preferred the more valuable option C over the more frequently rewarded option A in the low variance condition. With moderate variance, their choices became random, while with high variance, participants significantly favored the more frequently rewarded option A. Within the framework of our dual-process model, which juxtaposes frequency- and value-based systems, we identified a proportional increase in reliance on the frequency-based system as variance increased, which maps nicely onto participants' behavior. Model fitting further showed that the dual-process model significantly outperformed comparison models and accurately captured the observed preference changes. Together, these findings suggest that participants transition from a value-dominant strategy to a dual-process approach, with increasing reliance on reward frequency under conditions of high value uncertainty. This reliance reflects the use of reward frequency as a heuristic for value when estimating the underlying value distribution becomes increasingly challenging.

Related to our model, there have been some attempts to develop a dual-process framework. For example, Miller et al. (2019) juxtapose a habitual system with a goal-directed system, arbitrated by action-outcome contingency and habitization strength. While this framework is conceptually similar to our model, their habitual system relies solely on a tally of past choices without encoding the valence. Yet, later research[23] shows that frequency effects are not merely contingent on the act of selection, suggesting that valence may still play a role in the frequency-based system. In addition, without a distributional representation of the EV space, the measures of action-outcome contingency and habitization strength—somewhat analogous to the entropy measures in our value- and frequency-based systems—depend on separate, manually tuned calculations (e.g., habitization strength as the distance from the mean)[54]. This approach may lack the mathematical rigor and generalizability compared to the estimation of statistical dispersion through well-established distributional functions.

That said, the unique feature of this dual-process model lies in the entropy-based online weighting approach. This approach builds upon a rich behavioral modeling literature showing that decision-making is strongly influenced by uncertainty[55–57]. Higher uncertainty encourages exploration and learning[20,51], raising the "model temperature" during decision-making because individuals may deviate from a strictly value-based logic and consult alternative calculative mechanisms. However, most current efforts to incorporate uncertainty as a modulator of learning nevertheless rely on the Delta rule[20,58], and therefore, are inherently limited in their ability to explain behaviors that stem from alternative strategies, such as frequency effects. The challenge of formally defining and estimating uncertainty in a streamlined RL model without distributional representations often complicates its integration into a multi-system framework. Interestingly, one of the few studies that remotely addressed this issue in a model-based versus model-free framework used Dirichlet priors to estimate the level of uncertainty[59]. Moreover, uncertainty in one process does not necessarily translate into uncertainty across the entire environment. For instance, in our task, one might be unsure about the value difference between C and A (i.e., high Gaussian entropy) but confident that A has yielded much more cumulative rewards than C (i.e., low Dirichlet entropy). This could encourage individuals to switch decision-making strategies, driven by reduced cognitive efforts needed to ascertain the difference and risk-aversion. Therefore, the presented distribution-based dual-process model excels by recognizing how each process independently responds to the environment, with a unified and straightforward quantitative measure of uncertainty.

By introducing entropy as a weighting parameter for multiple parallel processes, we also implicitly introduced the idea that greater potential EV differences demand more rigorous validation. In our model, the frequency-based Dirichlet distribution is defined on a simplex, meaning that its components always sum up to one. This constraint reduces the potential statistical dispersion of the distribution but inherently limits the maximum distance between alternative options. In contrast, the multivariate Gaussian distribution is defined over the entire real space, allowing differences between options to be infinitely large but requiring a growing number of samples or a very small variance to validate the difference as reward values scale. The Dirichlet system can be seen as a distilled version of the Gaussian system. Since defining a "reward" inevitably requires input from the value-based system, reward encoding between the two systems is not guaranteed to operate entirely in parallel. Over time, however, detailed decision-making contexts (e.g., precise reward values) may fade into a simplified, offline tally of past outcomes classified in a binary way (i.e., rewarding/non-rewarding). This distinction is reflected in the difference in computational complexity between the Dirichlet and Gaussian distributions. This distillation process reduces the amount of information that needs to be retained and facilitates future retrieval of information stored in the Dirichlet system. This trade-off captures the balance between accuracy and cognitive efficiency. As the costs (e.g., cognitive effort, sample size) needed to validate decisions in the value-based system become prohibitively high, they may outweigh the potential gains from choosing a slightly better option, thereby encouraging a transition towards the simpler, frequency-based system.

Limitations

One limitation of the current dual-process model is that the subjective Dirichlet weight parameter (w_1) exhibited relatively lower stability during parameter recovery compared to the other two parameters. This may be because, in our paradigm, the subjective weight of the frequency-based system becomes less relevant when external value uncertainty is either extremely high or extremely low. In such cases, people are strongly biased toward either the frequency- or value-based system, regardless of their subjective preference. This is supported by the model-fitting results, which demonstrate that the purely objective dual-process model performed relatively well and successfully captured much of the dynamics involved in strategy switching. However, the inclusion of the subjective Dirichlet weight parameter still significantly improved model fit over the purely objective version, suggesting that this parameter captured meaningful variability in participant behavior for at least a subset of individuals whose preferences were not entirely determined by external uncertainty. This parameter may play a more critical role in future studies exploring individual differences in decision-making, especially when external uncertainty is held constant.

Conclusion

In conclusion, our findings of adaptive learning and decision-making in RL reconcile many findings of frequency effects in high variance (i.e., similar or higher than a binomial variance) experimental paradigms, such as the ABCD task[11,23], IGT[60–62], and Soochow Gambling Task[22,63], with other findings where decision-making appears to be primarily value-driven[64–69]. The dynamic interplay of reward variance, frequency, and uncertainty suggests that there is no definitive answer to which factor is more influential in RL, as it ultimately depends on the context. Methodologically, our model provides an exploratory approach to RL modeling by dissociating decision-making strategies rather than attempting to map out the entire cognitive pathway. The flexibility of such modeling framework not only allows for more accurate capturing of behavioral decision-making under various contexts, but also paves the way for future developments, such as adding biased priors or incorporating additional processes, which could broaden its potential to explain a wider range of human behaviors.

Data availability

Data presented in the current study can be accessed in Open Science Framework (https://osf.io/ks3nd/).

Code availability

Codes for statistical analyses and computational models mentioned in this study are available through Open Science Framework (https://osf.io/ks3nd/).

Received: 27 September 2024; Accepted: 4 April 2025;
Published online: 14 April 2025

References

1. Kahneman D. & Tversky A. Prospect Theory: An Analysis of Decision Under Risk. In; :99-127. 1979
2. Tversky, A. & Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992).
3. Ahn, W., Busemeyer, J. R., Wagenmakers, E. & Stout, J. C. Comparison of decision learning models using the generalization criterion method. *Cogn. Sci.* **32**, 1376–1402 (2008).
4. Plonsky, O., Teodorescu, K. & Erev, I. Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychol. Rev.* **122**, 621–647 (2015).
5. Erev, I. & Roth, A. E. Maximization, learning, and economic behavior. *Proc. Natl. Acad. Sci.* **111**, 10818–10825 (2014).
6. Gigerenzer, G. & Goldstein, D. G. Reasoning the fast and frugal way: Models of bounded rationality. *Psychol. Rev.* **103**, 650–669 (1996).
7. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annu Rev. Psychol.* **62**, 451–482 (2011).

8. Erev, I. et al. A choice prediction competition: Choices from experience and from description. *J. Behav. Decis. Mak.* **23**, 15–47 (2010).
9. Yechiam, E. & Busemeyer, J. R. Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychon. Bull. Rev.* **12**, 387–402 (2005).
10. Gonzalez, C. & Dutt, V. Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychol. Rev.* **118**, 523–551 (2011).
11. Don, H. J., Otto A. R., Cornwall A. C., Davis T. & Worthy D. A. Learning reward frequency over reward probability: A tale of two learning rules. *Cognition*. **193**, https://doi.org/10.1016/j.cognition.2019.104042 (2019).
12. Yon, D. & Frith, C. D. Precision and the Bayesian brain. *Curr. Biol.* **31**, R1026–R1032 (2021).
13. Colombo, M., Seriès P. Bayes in the brain—on Bayesian modelling in neuroscience. *Br J Philos Sci*. (2012). https://www.journals.uchicago.edu/doi/abs/10.1093/bjps/axr043?journalCode=bjps.
14. Mathys, C. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci*. **5**, https://doi.org/10.3389/fnhum.2011.00039 (2011).
15. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
16. Wulff, D. U., Mergenthaler-Canseco, M. & Hertwig, R. A meta-analytic review of two modes of learning and the description-experience gap. *Psychol. Bull.* **144**, 140–176 (2018).
17. Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G. & Palminteri, S. Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nat. Commun.* **9**, 4503 (2018).
18. Geisler, W. S. & Diehl, R. L. A Bayesian approach to the evolution of perceptual and cognitive systems. *Cogn. Sci.* **27**, 379–402 (2003).
19. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
20. 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.* **30**, 12366–12378 (2010).
21. Erev I., Roth A. E. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev*. 848-881 (1998). https://www.jstor.org/stable/117009?casa_token=3bWrhoc-xFgAAAAA%3A3dkfvvKZ5RSvw0hl-09GEiLDCby7ek9xWOzFjB2r9H11906uUXIY4QLQTFBFqGdetqwBUwSDSuVXyGwtJX86J7Jz0oez4we9atBhsJ1oPi-tVVllsMqu.
22. Chiu, Y. C. et al. Immediate gain is long-term loss: Are there foresighted decision makers in the Iowa Gambling Task? *Behav. Brain Funct.* **4**, 13 (2008).
23. Don, H. J. & Worthy, D. A. Frequency effects in action versus value learning. *J. Exp. Psychol. Learn Mem. Cogn.* **48**, 1311–1327 (2022).
24. Niv, Y., Edlund, J. A., Dayan, P. & O'Doherty, J. P. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* **32**, 551–562 (2012).
25. Levy, H. & Markowitz, H. M. Approximating expected utility by a function of mean and variance. *Am. Econ. Rev.* **69**, 308–317 (1979).
26. Dezfouli, A., Griffiths, K., Ramos, F., Dayan, P. & Balleine, B. W. Models that learn how humans learn: The case of decision-making and its disorders. *PLoS Comput Biol.* **15**, e1006903 (2019).
27. Yechiam, E. Acceptable losses: the debatable origins of loss aversion. *Psychol. Res.* **83**, 1327–1339 (2019).
28. Ferguson, T. S. A Bayesian analysis of some nonparametric problems. *Annals Stat*. 209–230 (1973). https://www.jstor.org/stable/2958008?casa_token=jY6oA67PuwMAAAAA%3Aj4YQzVCmRZSF3SFes0mac2cn_UNu61TJsecgax7gaKu3GTlU_2PC0V4dLM6qOFru4Yk32X_oSQQnicgJOEo5k9sWhVgMn3KmPTAOJDJLh_eOwhgFC_2x
29. Navarro, D. J., Griffiths, T. L., Steyvers, M. & Lee, M. D. Modeling individual differences using Dirichlet processes. *J. Math. Psychol.* **50**, 101–122 (2006).
30. Wunderlich, K., Smittenaar, P. & Dolan, R. J. Dopamine enhances model-based over model-free choice behavior. *Neuron* **75**, 418–424 (2012).
31. Eppinger B. & Walter M., Heekeren H. R., Li S. C. Of goals and habits: age-related and individual differences in goal-directed decision-making. *Front. Neurosci*. **7**, https://doi.org/10.3389/fnins.2013.00253 (2013).
32. Eppinger, B., Walter, M. & Li, S. C. Electrophysiological correlates reflect the integration of model-based and model-free decision information. *Cogn. Affect Behav. Neurosci.* **17**, 406–421 (2017).
33. Molinaro, G. & Collins, A. G. E. Intrinsic rewards explain context-sensitive valuation in reinforcement learning. *PLoS Biol.* **21**, e3002201 (2023).
34. Cover T. M., Thomas J. A. Differential Entropy. In: *Elements of Information Theory*. Wiley; 2005:243-259.
35. Yechiam, E. & Ert, E. Evaluating the reliance on past choices in adaptive learning models. *J. Math. Psychol.* **51**, 75–84 (2007).
36. Collins, A. G. E. & Shenhav, A. Advances in modeling learning and decision-making in neuroscience. *Neuropsychopharmacology* **47**, 104–118 (2022).
37. Sutton, R. S. & Barto, A. G. Reinforcement learning. *J. Cogn. Neurosci.* **11**, 126–134 (1999).
38. Anderson, J. R. ACT: A simple theory of complex cognition. *Am. Psychol.* **51**, 355 (1996).
39. Anderson, J. R., Matessa, M. & Lebiere, C. ACT-R: A theory of higher level cognition and its relation to visual attention. *Hum. Comput Interact.* **12**, 439–462 (1997).
40. Kool, W., Cushman, F. A. & Gershman, S. J. When does model-based control pay off? *PLoS Comput Biol.* **12**, e1005090 (2016).
41. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* **19**, 716–723 (1974).
42. Schwarz G. Estimating the dimension of a model. *Annals Stat*. 461–464 (1978). https://ieeexplore.ieee.org/abstract/document/1100705?casa_token=_06a4Or_4pcAAAAA:Zwz2OpH5buldIC1YrK6yQ_k6gNO9_-kbTAxQLd78aAgC1Zw-3Rj2RgNb8pmXK_Zu3P7NmPoQ7g.
43. Wagenmakers, E. J. & Farrell, S. AIC model selection using Akaike weights. *Psychon. Bull. Rev.* **11**, 192–196 (2004).
44. Wagenmakers, E. J. A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.* **14**, 779–804 (2007).
45. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *Neuroimage* **46**, 1004–1017 (2009).
46. Wilson, R. C. & Collins A. G. Ten simple rules for the computational modeling of behavioral data. *Elife*. **8**, https://doi.org/10.7554/eLife.49547 (2019).
47. Cohen, J. D., McClure, S. M. & Yu, A. J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. B: Biol. Sci.* **362**, 933–942 (2007).
48. Speekenbrink, M. Chasing unknown bandits: Uncertainty guidance in learning and decision making. *Curr. Dir. Psychol. Sci.* **31**, 419–427 (2022).
49. Mukherjee, K. A dual system model of preferences under risk. *Psychol. Rev.* **117**, 243–255 (2010).
50. Diederich, A. & Trueblood, J. S. A dynamic dual process model of risky decision making. *Psychol. Rev.* **125**, 270–292 (2018).
51. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
52. Jansen, R. A., Rafferty, A. N. & Griffiths, T. L. A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. *Nat. Hum. Behav.* **5**, 756–763 (2021).

53. Melnikoff, D. E. & Strohminger, N. Bayesianism and wishful thinking are compatible. *Nat. Hum. Behav.* **8**, 692–701 (2024).
54. Miller, K. J., Shenhav, A. & Ludvig, E. A. Habits without values. *Psychol. Rev.* **126**, 292–311 (2019).
55. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
56. Yu, A. J. & Dayan, P. Uncertainty, Neuromodulation, and Attention. *Neuron* **46**, 681–692 (2005).
57. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
58. Bach, D. R. & Dolan, R. J. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat. Rev. Neurosci.* **13**, 572–586 (2012).
59. Lee, S. W., Shimojo, S. & O'Doherty, J. P. Neural computations underlying arbitration between model-based and model-free learning. *Neuron* **81**, 687–699 (2014).
60. Lin, C. H., Chiu, Y. C., Lee, P. L. & Hsieh, J. C. Is deck B a disadvantageous deck in the Iowa Gambling Task? *Behav. Brain Funct.* **3**, 16 (2007).
61. Toplak, M. E., Jain, U. & Tannock, R. Executive and motivational processes in adolescents with Attention-Deficit-Hyperactivity Disorder (ADHD). *Behav. Brain Funct.* **1**, 8 (2005).
62. Wilder, K. E., Weinberger, D. R. & Goldberg, T. E. Operant conditioning and the orbitofrontal cortex in schizophrenic patients: unexpected evidence for intact functioning. *Schizophr. Res.* **30**, 169–174 (1998).
63. Lin, C. H., Chiu, Y. C. & Huang, J. T. Gain-loss frequency and final outcome in the Soochow Gambling Task: A reassessment. *Behav. Brain Funct.* **5**, 1–9 (2009).
64. Busemeyer, J. R. & Myung, I. J. An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *J. Exp. Psychol. Gen.* **121**, 177 (1992).
65. Lee, S., Gold, J. I. & Kable, J. W. The human as delta-rule learner. *Decision* **7**, 55–66 (2020).
66. Ritz, H., Nassar, M. R., Frank, M. J. & Shenhav, A. A control theoretic model of adaptive learning in dynamic environments. *J. Cogn. Neurosci.* **30**, 1405–1421 (2018).
67. Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J. & Lamm, C. Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. *Soc. Cogn. Affect Neurosci.* **15**, 695–707 (2020).
68. Welford, B. P. Note on a method for calculating corrected sums of squares and products. *Technometrics* **4**, 419–420 (1962).
69. Ling, R. F. Comparison of several algorithms for computing sample means and variances. *J. Am. Stat. Assoc.* **69**, 859–866 (1974).

Author contributions

M.H.: Conceptualization, Methodology, Validation, Formal analysis, Writing—Original Draft, Writing—Review & Editing, Visualization. H.J.D.: Conceptualization, Methodology, Investigation, Data Collection, Writing—Review & Editing. D.A.W.: Conceptualization, Resources, Writing—Review & Editing, Supervision, Project administration.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44271-025-00249-y.

Correspondence and requests for materials should be addressed to Mianzhi Hu.

Peer review information *Communications Psychology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Erdem Pulcu and Jennifer Bellingtier. [A peer review file is available.]