A Foraging-Theory Based Model Captures The Full Spectrum of Human Behavioral Diversity in a Classic RL Task

Dameon C Harrell[1], Meriam Zid[2], Veldon-James Laurie[2], Cathy S Chen[1], Nicola M Grissom[1], David P Darrow [1], R Becket Ebitz[2◕], Alexander B Herman[1*◕]

1 Department of Psychiatry, University of Minnesota, Minneapolis, MN, 55455, USA
2 Department of Neuroscience, University of Montreal, Montreal, QC , H3T 1J4, Canada.

◕These authors contributed equally to this work.
* correspondingauthor@institute.edu

Abstract

Decision-making tasks involving multiple, simultaneously presented options are mainstays of cognitive neuroscience and psychology, and are increasingly important to the emerging field of computational psychiatry. Modeling approaches to these tasks overwhelmingly assume that participants make choices based on explicitly comparing the values of the presented options. Contrary to this long-held assumption, we found instead that humans employ a compare-to-threshold decision process, similar to theories of foraging, when making sequential decisions about concurrently available options. We confirmed this result in a large (1000 participant) dataset with multiple converging lines of evidence comparing both model fits and model generative performance. Value-comparison models were restricted to a reduced area of the potential space of single-trial outcome-dependent behavior, demonstrating an intrinsic limitation in the ability to reproduce strategy diversity. Furthermore, we found that using even the best-fit value-comparison model led to a substantial, systematic bias and a compression of individual differences in reconstructed behavior compared to the foraging-based model, leading to weaker predictions of behavioral health measures. Our results imply that studies using value-comparison models to link behavior with neural activity or psychiatric symptoms may be less sensitive to individual differences than a simple alternative based on ethological foraging.

Author summary

This study challenges the core assumption of reinforcement learning (RL) models in sequential decision-making: that choices are made by comparing option values. We demonstrate that most individuals actually employ a foraging-like mechanism, where decisions to explore or exploit depend on comparing exploitation value against a threshold. The foraging model not only achieves better conventional fitness metrics (AIC, BIC) but also captures the full spectrum of reward responsiveness observed in human data, while traditional RL models represent only a restricted subset. This limitation causes RL models to compress individual differences and introduce systematic bias. Furthermore, the foraging-like model explains mental health symptoms more effectively and parsimoniously than traditional RL approaches. Our findings suggest that sequential decision-making should be reconsidered through the lens of foraging

theory, potentially enhancing our understanding of both cognitive mechanisms and their relationship to psychopathology.

Introduction

Adaptive decision-making—balancing exploration of new opportunities with exploitation of known options—is fundamental to human behavior and frequently disrupted in psychiatric disorders [1]. Tasks probing these processes offer powerful tools for identifying individual differences in decision-making strategies. Exploration-exploitation impairments are consistently observed in conditions including depression, schizophrenia, autism, and chronic stress, making them valuable targets for computational psychiatry approaches that link computational-model parameters to psychiatric symptoms.

A critical challenge in this field is selecting tasks and models that accurately capture the full spectrum of human behavior—a prerequisite for understanding neurodivergent or pathological deviations. In this study, we examine the "restless bandit" task, which simulates dynamic reward environments to elicit realistic exploration-exploitation behavior. Most models of this task, for instance, reinforcement learning models, assume that people make their choices by tracking and comparing the values of all available options [2]. Although sophisticated formulations of these models have been developed to expand and improve their functioning, the foundational assumption of value-comparison based-choice is rarely questioned.

We recently developed an alternative "compare-to-threshold" model based on foraging theory [3], which assumes instead that individuals maintain a focal option and compare its value to an internal threshold to decide whether to exploit or explore. This approach offers computational efficiency (tracking only a single value and threshold regardless of option count) and conceptual alignment with psychiatric constructs through its threshold parameter, which reflects an individual's perception of environmental reward potential. While similar models have been applied to patch-foraging tasks [4], they have not been applied to bandit tasks.

In a large sample (n=1000) performing a three-armed restless bandit task, we found that our foraging-based threshold model outperformed value-comparison approaches in explaining both the breadth and depth of human behavior. Furthermore, the foraging-based model's parameters better captured psychiatric symptom variability than parameters from the best-fitting value-comparison models, providing additional evidence that a compare-to-threshold model aligns more closely with the underlying cognitive process.

Results

Task and Participant Behavior

We recruited 1000 gender-matched adults (500 males, 500 females) to participate in a restless 3-arm bandit task (Fig 1a). Participants selected one of three playing cards by moving their mouse over their choice. Outcomes were binary: participants either received a point (win) or no point (loss). The probability of winning associated with each deck randomly drifted over time, encouraging participants to balance exploitation of good options with exploration for potentially better alternatives. Performance was evaluated by comparing the total number of rewarded trials to chance expectations. The vast majority of participants (925/1000) accumulated more rewards than would be expected by chance (see Methods for statistical model details).

Fig 1. Task design, model comparison, and behavioral fits. (A) Three-deck bandit task interface. (B) Random-walk reward probabilities for each deck over time. (C) CTT RL model diagram (top) and switching probabilities at $\beta = 2$ and $\beta = 10$ (bottom), showing threshold-dependent behavior. (D) Simple RL model diagram (top) and switching probabilities as a function of value difference (bottom), illustrating greediness effects. (E) Model comparison using AIC/BIC weights shows CTT RL outperforming all value-comparison models: $\sim 60\%$ vs $< 18\%$ probability of being correct for the best value-comparison model). (F) Model recovery analysis showing limited overlap between CTT and CK RL generated data fits, indicating distinct decision processes. (G) Actual vs. predicted lose-switch rates showing CTT RL accurately predicts human behavior across the range, while CK RL fails above ~ 0.7. (H) Average switch percentages following wins/losses for human participants and model simulations, demonstrating significant differences between human/CTT data and CK RL predictions.

Model Comparisons

To determine whether participants' decision strategies more closely resembled a compare-to-threshold (CTT) reinforcement learning model or one of several value-comparison RL models, we fitted each participant's choice sequence using maximum likelihood estimation (MLE) [5]. We calculated Akaike's Information Criterion (AIC) weights [6] and Bayesian Information Criterion (BIC) weights for model comparison. These weights reflect the probability that a particular model better explains the observed choices than all other candidate models [7], with BIC being more conservative for complex models by accounting for the number of observations.

The CTT RL model provided the best fit for most participants—578/1000 and 606/1000 according to AIC and BIC weights, respectively. In contrast, the best-fitting value-comparison model (Choice Kernel RL or CK RL) provided the best fit for only 181/1000 (AIC) and 136/1000 (BIC) participants (Table 1, Fig 1e). While the DLR (dual learning rate) RL model produced a slightly higher overall BIC weight (by 0.01) than the CK RL model, we selected the CK RL model to represent the value-comparison models for further analysis based on its larger AIC weight. This focus on one value-comparison model is justified since all value-comparison models share the same fundamental limitation compared to the CTT RL model.

Table 1. Model Fits Comparison

Model	Raw AIC	w(AIC)	n(AIC)	Raw BIC	w(BIC)	n(BIC)
DLR RL	204.7 (80)	0.11 (0.23)	113	215.8 (80)	0.14 (0.27)	150
Noisy RL	205.1 (76.5)	0.05 (0.15)	45	219.9 (76.5)	0.03 (0.11)	22
CK RL	199.6 (76.3)	0.18 (0.33)	181	214.5 (76.3)	0.13 (0.3)	136
EB RL	202.9 (83)	0.09 (0.22)	83	214 (83)	0.11 (0.23)	86
CTT RL	185.3 (82.7)	0.56 (0.45)	578	196.4 (82.7)	0.59 (0.45)	606

Average (across participants) raw AIC, AIC weights, raw BIC, and BIC weights. The table also contains the number of participants for whom the model was the best fit according to both AIC and BIC weights. Standard deviations are in parentheses.

Model Recovery Analysis

The model comparison results above rely on predictive performance (how well a model predicts observed choices) and parsimony (favoring models with fewer parameters) [8].

However, these criteria only make relative comparisons—they identify which model is best among candidates without assessing absolute model quality. Additionally, models with equal parameter counts can differ in their flexibility to fit data. To address these concerns, we performed model recovery analysis to evaluate how well the models could mimic each other [8–10].

We used the Parametric Bootstrap Cross-Fitting Method (PBCM) to quantify each model's ability to account for data generated by an alternative model [9]. When models cannot effectively mimic each other, this is evidence they have distinct decision-making processes, and we can more confidently interpret differences between AIC and BIC weights. Fig 1f displays the distributions of differences in goodness-of-fit $\Delta gof = gof_{CK} - gof_{CTT}$ between simulated data from the CTT RL and CK RL models. Negative Δgof values indicate the CK RL model fit better, while positive values indicate the CTT RL model fit better.

For data generated by the CTT RL model (green bars), approximately 13.5% of simulations were better fit by the CK RL model. Conversely, for data generated by the CK RL model (gold bars), approximately 5% were better fit by the CTT RL model. The limited overlap between these distributions suggests the decision processes of the CK and CTT RL models are distinct, supporting the reliability of our AIC and BIC weight results favoring the CTT RL model.

Generative Performance

While predictive performance provides valuable model comparison information, it is insufficient for falsifying particular models or measuring their capacity to produce behaviors of interest. Therefore, a model's "generative performance"—its ability to reproduce key behavioral patterns—is also important to consider [8].

Fig 1g demonstrates the generative performance of both CTT and CK RL models in reproducing human switching behavior following losses in the bandit task. Each point (green for CTT, gold for CK) represents the average lose-switch rate across 50 simulations (300 trials each) using parameters optimized for individual participants. The results clearly show that the CTT RL model more accurately reproduces human switching behavior following losses compared to the CK RL model.

Value-Comparison Model Dynamics and Limitations

To understand why value-comparison models struggle to reproduce switching behavior after losses, we analyzed the dynamics of the softmax decision rule across different value states with learning rate, α, set to 1. At this learning rate, an option's value is immediately updated to match the most recent outcome (1 for win, 0 for loss), resulting in eight possible value states (2^3) combinations for three options.

Fig 2a presents state transition diagrams for value-comparison (w/ basic softmax) RL models with different inverse temperature (β) values. In these diagrams, V_{cc} represents the current choice's value, followed by the values of the two alternative options.

The left diagram ($\beta = 8$) shows "greedy" decision-making where the model strongly prefers higher-valued options. When the alternative options have zero value (state: V_{cc}, 0, 0), the probability of switching after a win is zero. Switching after a win (win-switch) only occurs when at least one alternative option has a value of one. After a loss, the switching probability is 100%, when an at least one alternative option's value is one (due to high β). But, once the model reaches state (V_{cc}, 0, 0), it cannot return to states where alternatives have non-zero values because it will not switch after a win. This limits the maximum lose-switch probability to 67%.

The right diagram ($\beta = 2$) shows less deterministic decision-making. Here, the win-switch probability in state (V_{cc}, 0, 0) increases to 21%, allowing the model to

Fig 2. Value-Comparison Model Limitations in Outcome Sensitivity. (A) Value-state transition diagrams for softmax decision rule ($\alpha = 1$) at two different β values. With $\beta = 8$ (left), "greedy" decisions trap the model in state "$V_{cc}, 0, 0$" limiting lose-switch probability to 67%. At $\beta = 2$ (right), increased stochasticity allows occasional win-switches, enabling higher lose-switch rates. (B) Outcome sensitivity space with win-switch rates (vertical) and lose-switch rates (horizontal). The shaded region below the diagonal represents lose-switch % >= win-switch % behavior; corners represent extreme sensitivities/insensitivities. (C) Human participants' outcome sensitivity (gray dots) with behavioral ranges producible by CTT RL (green outline) and CK RL (gold outline) models. (D) Best-fit parameter simulations for CK RL (gold) and CTT RL (green) models. CTT simulations match human distribution, while CK simulations show compressed sensitivity range and systematically lower lose-switch rates. (E) Polar histograms of difference vectors between participant and model-simulated sensitivities, showing CTT RL's evenly distributed errors versus CK RL's systematic bias toward underestimating lose-switch rates.

occasionally reach states where alternative options have non-zero values after losses.

In summary, the softmax decision rule creates an inherent trade-off: higher lose-switch rates require higher win-switch rates, which depend on β . When β is too high, lose-switch rate is capped at 67%. When $\beta = 0$ (completely random), the lose-switch rate is also 67%. The maximum lose-switch rate occurs at an intermediate β value but remains below 100% due to the required stochasticity.

Outcome Sensitivty Framework

We conceptualize these limitations using "outcome sensitivity"—how responsive an agent is to wins or losses. We define a two-dimensional outcome sensitivity space with lose-switch rates on one axis and win-switch rates on the other (Fig 2b).

This space has several important reference points: maximal outcome sensitivity (lose-switch = 100%, win-switch = 0%, also known as a win-stay/lose-switch strategy) appears in the lower right corner; maximally biased insensitivity (either never switching or always switching regardless of outcome) appears in the lower-left and upper-right corners; and unbiased insensitivity (random 50% switching probability) appears at the midpoint (0.5, 0.5). The area below the diagonal (lose-switch > win-switch) is most relevant for reward-maximizing behavior. The lower right corner has particular relevance as an anchor point because a pure win-stay/lose-switch (also known as Pavlov) strategy is a simple heuristic solution to a wide range of decision-making problems ([11]; [12]; [13]), including bandits ([14]). An individual point in this space thus represents the degree to which that participant's behavior deviates from win-stay/lose switch along dimensions of sensitivity to loss and sensitivity to wins.

Human and Model Behavior in Outcome Sensitivity Space

Fig 2c displays human participants' behavior in outcome sensitivity space (gray dots). The vast majority (98.3%) showed higher lose-switch than win-switch rates, with 64.8% having win-switch rates below 10%. Over a quarter (252/1000) had win-switch rates below 10% and lose-switch rates above 80%.

The gold and green outlines in Fig 2c show the behavioral ranges the CK RL and CTT RL models can produce. The CTT RL model covers the entire region where lose-switch rates exceed win-switch rates. In contrast, the CK RL model covers only a portion of this space. Its maximum win-switch rate is 67% (with completely random

selection, $\beta = 0$). When win-switch is 0%, the lose-switch rate is limited to 67%. Higher lose-switch rates are only possible with increased win-switch rates.

The CK RL model, particularly its softmax decision rule, cannot produce the full range of lose-switch behavior because the lose-switch rate depends partly on the win-switch rate. Additionally, maximum lose-switch and win-switch rates depend on the number of available options.

Fig 2e? uses polar histograms to visualize differences between human and model-simulated behavior. Specifically, it displays the distributions of the angles of the vectors formed by the line segments connecting each human participant's point in outcome sensitivity space to the point of the corresponding best-fit optimized CTT RL and CK RL simulations. The angles of CTT RL difference vectors show a relatively even distribution around the polar plane, indicating no systematic bias. In contrast, the CK RL difference vector angles cluster between 165-195 degrees, with 673 out of 1000 simulations underestimating participants' lose-switch rates more than they misestimate win-switch rates.

The magnitude of these vectors further confirms CTT RL's superior generative performance. Most CTT RL simulations (695/1000) are within 0.02 switches/outcome of human data, with none exceeding 0.18. CK RL simulations show a more uniform distribution of differences, with some as large as 0.65 switches/outcome, and nearly half (484/1000) exceeding the maximum CTT RL difference. The magnitude is calculated as the euclidean distance between the true and simulated outcome sensitivity points (Eq (1)).

$$distance = \sqrt{(ls_{human} - ls_{model})^2 + (ws_{human} - ws_{model})^2} \tag{1}$$

Outcome Sensitivity in Model Parameter Space

We defined a unified outcome sensitivity measure as: (lose-switch rate - 0.5) + (win-stay rate - 0.5), where win-stay = 1 - win-switch. This yields: maximum sensitivity at lose-switch = 1, win-stay = 1 (outcome sensitivity = 1); unbiased insensitivity at lose-switch = 0.5, win-stay = 0.5 (outcome sensitivity = 0); and maximum biased insensitivity when one rate is 1 and the other is 0 (also outcome sensitivity = 0). We then examined this measure across the different models we considered.

Simple RL Model

For the simple value-comparison RL model (Fig 3a), when $\beta > 2$, outcome sensitivity primarily depends on α, showing a clear gradient from high to low sensitivity as α decreases. When $\beta < 2$, outcome sensitivity mainly depends on β and is relatively independent of α. Outcome sensitivity reaches a maximum of approximately 0.68 and is minimized when either β or α equals zero.

Fig 3

CK RL Model

The CK RL model (Fig 3b) shows a more complex parameter landscape with no clear gradient in outcome sensitivity relative to α and β . Low outcome sensitivities (< 0.2) appear throughout the parameter space, even with high α and β values. The highest sensitivities occur with high α and β, but the choice kernel parameters (α_{ck} and β_{ck}) often drive behavior toward biased insensitivity.

Fig 3c shows that high β_{ck} values (> 5) generate consistently low outcome sensitivities (< 0.2) by biasing behavior toward repeatedly choosing the same option regardless of outcome. Outcome sensitivity appears inversely proportional to β_{ck} when $\beta_{ck} < 3$, but shows little dependency on α_{ck}.

CTT RL Model

The CTT RL model's outcome sensitivity (Fig 3d) resembles the Simple RL model with important differences. When $\beta > 2$, outcome sensitivity primarily depends on α, but unlike Simple RL, the relationship isn't monotonic as α decreases. Outcome sensitivity varies from 1 to 0.4 even with high $\beta(> 6)$ and high α, $\sim (0.8 - 1)$. CTT RL can produce biased outcome insensitivity with α as high as 0.7, while Simple RL requires $\alpha < 0.2$.

Fig 3e shows outcome sensitivity as a function of α and threshold parameter ρ. Outcome sensitivity is roughly proportional to α and symmetric about $\rho = 0.5$, decreasing as ρ moves away from 0.5 in either direction. The range of possible sensitivities narrows as α decreases, forming a triangular pattern in α-ρ space.

The threshold ρ functions as a bias term that decouples sensitivity to wins from sensitivity to losses. This allows the model to achieve maximum sensitivity to both simultaneously, which is impossible for value-comparison models using the softmax decision process, where sensitivity to losses is inherently tied to sensitivity to wins.

Prediction of behavioral measures

We hypothesized that the CTT's superior ability to capture individual differences in task behavior would extend to better predictions of personality and mental health symptom measures. We analyzed a range of such self-report measures, covering depression, anxiety, pain, sleep quality, anhedonia, motivation, autism-like traits, and cognitive flexibility. Models predicting these measures from the CTT parameters outperformed CK RL models on 10/13 of these scales by AIC weight. We next examined in depth two of these domains, apathy and anxiety, which we have recently shown to co-vary with explore-exploit behavior on this task. The CTT parameter-based model better accounted for variance in both symptom domains. The CTT parameters also provided a more parsimonious description of the cognitive processes that are related to apathy and anxiety. For both symptoms, the explained variance was concentrated in one CTT parameter, the foraging threshold ρ, while it was more evenly distributed across the parameters of the CK RL model. Furthermore, the single CTT ρ parameter explained more variance in both of these symptoms measures than all of the CK RL parameters combined.

Fig 4. *Comparing the predictive utility of CK RL and CTT RL models for anxiety, depression, and other self-reported measures* A) Proportion of variance (R^2) in generalized anxiety (left) and depressive (right) symptoms explained by RLCK and Foraging models. For each model, the total variance explained is further broken down into the top contributing parameter from each model (α for RLCK; and ρ for Foraging). B) Normalized differences in R^2 across a range of self-reported measures. Bars above the zero threshold indicates that the Foraging model explains more variance in that measure; bars below the zero threshold indicates that the RLCK model explains more variance.

Discussion

Our findings provide evidence that humans utilize a compare-to-threshold (CTT) decision process rather than a value-comparison approach when making sequential decisions about concurrently available options. This represents a fundamental shift in our understanding of human decision-making, with significant implications for computational psychiatry and cognitive neuroscience.

The superior performance of the CTT RL model over traditional value-comparison models in explaining human behavior stems from its ability to capture a much broader range of decision strategies. The limitations of value-comparison models are not merely implementation details but reflect fundamental constraints in their architecture. Specifically, the softmax decision rule used in these models creates an intrinsic trade-off between win-switch and lose-switch rates, leading to model value-states with bounded lose-switch or win-switch rates. These bounds prevent value comparison models from populating the full outcome-sensitivity lose-switch ¿ win-switch subspace, restricting them to a smaller sub-region of this subspace instead. As a result, when fit to real human behavior, value-comparison models compress the observed behavior into this sub-region, obscuring individual differences and introducing systematic biases. This compression implies that the parameters from even the best-fit value-comparison model must be interpreted with caution. In contrast, the CTT model's threshold parameter allows for independent modulation of win-switch and lose-switch rates, enabling it to capture the full range of possible strategies and thus reproduce the full range of observed human behavior.

Our findings have important implications for computational psychiatry, underscoring the importance of accurate modeling of individual differences [15]. The constrained outcome-sensitivity and systematic bias towards lower lose-switch rates introduced by value-comparison models compresses individual differences and may obscure or even misconstrue important relationships between decision-making parameters and psychiatric symptoms. The inability of these models to represent agents with high lose-switch rates and low win-switch rates—a pattern we observed in many human participants—could lead researchers to mischaracterize psychiatric populations or miss important correlations between model parameters and clinical symptoms. In contrast, the coverage of outcome-sensitivity space produced by the CTT model resulted in a superior ability to capture individual differences in a diverse range of symptom measures. In two important symptom domains, apathy and anxiety, a single parameter from the CTT model (the threshold) explained more variance than all of the parameters combined from the best-fitting value comparison model. This parsimony offers significant advantages for inferring the connections between symptoms and cognitive process and, ultimately, putative physiological mechanisms [16].

The threshold parameter in the CTT model may have special relevance for psychiatric research as it may relate to an individual's perception of environmental reward, as it does in classic foraging theory ([3]; [17]). Altered threshold values could represent different expectations about reward availability—a construct directly relevant to various neuropsychiatric disorders where reward expectation is affected. For instance, we found that thresholds decrease as apathy increases, implying that people's overall level of motivation is related to how rewarding they see the environment. In contrast, thresholds increased with anxiety, potentially reflecting a response to risk aversion or the worry that unchosen options may be better than the choice they've made. Future research should explore whether threshold differences can differentiate clinical populations and whether these parameters are more sensitive to treatment effects than those derived from value-comparison models.

Several limitations of the present study should be addressed in future work. First, while our task uses a "restless bandit" paradigm to encourage exploration-exploitation

decisions, other decision-making contexts might elicit different strategies. Second, our CTT model utilized a constant threshold, but implementing a dynamic threshold that updates based on reward history could further improve model fit and better represent adaptation to changing environments. Third, although we examined a wide range of reinforcement learning models, there may be other possible formulations that would overcome the limitations that seem to be inherent to value comparison.

It is worth noting that value-comparison models have been foundational in computational neuroscience, leading to important insights about dopaminergic prediction error signals and reward learning ([18]; [19]). Because prediction error-based learning is central to both value-comparison and compare-to-threshold models, our results suggest that some of these past findings may warrant re-examination from the perspective of foraging theory. In our foraging-based model, the prediction refers to the value of the decision, stay or switch, rather than the values of the choices. This distinction may yield new insights into how value signals and thresholds are represented in the brain.

Our results challenge a fundamental assumption in decision-making research and offer a more ecologically valid framework for understanding how humans navigate exploration-exploitation dilemmas. The compare-to-threshold perspective provides superior explanatory power for individual differences in decision strategies and may ultimately advance our understanding of decision-making impairments in psychiatric disorders.

Methods

Participants

The required number of participants was determined through a priori power analysis. To detect correlations of r = 0.1 with symptom measures (typical for individual differences research) with 80% power at $\alpha = 0.05$, we required a minimum sample of 782 participants. We recruited 1500 participants to account for expected exclusions based on previous large-sample online studies ([20], [21]) and our own pilot work, expecting to achieve a sample of between 900-1100 participants, thus allowing for a buffer above the minimum sample size. Our final sample of 1000 participants (500 female) provided 98% power to detect r = 0.1. We recruited a sample of 1512 participants (non-clinical sample) via Prolific (Prolific. co); exclusion criteria included current or history of neurological and psychiatric disorders. Participants were excluded if they did not complete all questionnaires (3.57% of initial sample) or they did not complete the bandit task (30.22% of initial sample) (Table S1). 1000 participants completed all questionnaires and the restless bandit task (age range 18-54, mean ± SD = 28.446 ± 10.354 years; gender, 493 female). All participants were compensated for their time in accordance with Prolific's minimum wage.

Task

Each participant completed a 300, 3-arm restless bandit task. On each trial, participants were presented with images of 3 face-down playing cards, each associated with a deck of a specific color(arm), and instructed to choose one by moving the cursor over their choice. Following every choice, participants received feedback: either 1 point (a win) or no points (a loss). The probability of winning on any trial was independent for each deck and the product of a "random walk". There was a 67% probability that on any trial a given deck might increase or decrease its win yield rate by an increment of 10%.

Raw Performance Assessment

The number of wins expected by chance for each participant's reward schedule was determined by generating distributions (probability mass functions; PMF) of the number of winning trials accumulated by randomly selecting an arm on each of 300 trials for that participant's reward schedule. Specifically, 1000 300-trial simulations were used to produce each participant's PMF for chance expected number of wins. The number of winning trials actually accumulated by each participant was then used to determine the probability of accumulating that number of wins by chance according to the PMF. If the probability was less than .05, the participant's number of wins was assessed as being non-chance.

Model Details

Compare-to-Threshold (CTT) (foraging based) RL

CTT RL is a model comprising a learning rule Eq (2), for estimating option values, and a decision rule. In contrast to the more common version of the delta learning rule (see Eq (5), only the value of the currently selected option is maintained and the value resets to .5 after a decision to switch to an alternative option. Learning rate $\alpha(\in [0,1])$ determines the rate at which current outcome, $0(t)$, is incorporated into value estimate, $v(t)$.

$$v(t+1) = \begin{cases} v(t) + \alpha \cdot (O(t) - v(t)) & \text{after stay} \\ .5 & \text{after switch} \end{cases} \tag{2}$$

The CTT RL model uses a compare-to-threshold decision rule, Eq (3). The single maintained value is compared, via a difference in the exponential, to a threshold, $\rho(\in [0,1])$, a parameter of the model. $\beta(\in \mathbf{R}^+)$, another model parameter, adjust sensitivity to the difference between $v(t)$ and ρ, thus partially determining stochasticity in the decision process. Our CTT decision rule outputs the probability of switching, not the probability of choosing a *particular* option. When the model decides to switch, it chooses between the previously unselected alternatives with equal probability.

$$P(switch) = \frac{1}{1 + e^{\beta \cdot (v(t) - \rho)}} \tag{3}$$

Softmax (value-comparison) Decision Rule

The primary difference between the foraging based CTT RL and the value-comparison RL models is found in their decision rules. All of the value-comparison models in our study use the softmax equation Eq (4), aka the Boltzmann distribution [22], either in its basic (DLR RL) or a modified form (CK, EB, and Noisy RL). The output of the softmax is the probability of choosing a particular option. More precisely, the softmax models decisions of "non-greedy" choices based on the relative values of the available options. Here, greedy is defined as always selecting the option with the highest value. The right-hand side of the top portion of Eq (4) can be rearranged to see the value comparisons (differences) more explicitly. This rearrangement is shown in the bottom half of Eq (4). $P[C(t,i) = 1]$ is the probability of choosing option i on trial t; $C(t,i) = 1$ if option i was selected on trial t, $C(t,i) = 0$ otherwise. $\beta(\in \mathbf{R}^+)$ (aka "inverse temperature") is commonly thought of as controlling the amount of noise in the decision process and therefore determines its stochasticity; β also determines the sensitivity to value differences. When β is zero, the decision process is insensitive to value difference, completely random, and chooses uniformly between each option. As β approaches

positive infinity, the softmax output becomes "greedy": it always selects the (an) option with the highest value.

$$
\begin{aligned}
P[C(t,i)=1] &= \frac{e^{\beta \cdot v_i(t)}}{\sum_{j=1}^{N} e^{\beta \cdot v_j(t)}} \\
&= \frac{1}{1+\sum_{j=1,j!=i}^{N} e^{\beta \cdot (v_j(t)-v_i(t))}}
\end{aligned}
\tag{4}
$$

Choice Kernel (CK) RL

The CK RL model [5] employs two learning rules and uses a modified softmax for its decision rule. The basic delta learning rule, Eq (5) [22,23], updates the value of options. Learning rate $\alpha(\in [0,1])$ determines the rate at which the current outcome, $0(t)$, is incorporated into value estimate, $v_j(t)$. $O(t)=1$ for a win or $O(t)=0$ for a loss. Only the value of chosen option is updated during a trial; $\delta_j(t)=1$ if option j was chosen during trial t and is 0 otherwise.

An additional learning rule is used to maintain an estimate of how frequently each option is chosen in order to account for the tendency of individuals to repeat their previous actions [5]. Choice frequency estimates are updated on each trial according to Eq (6). $ck_j(t)$ is the estimated choice frequency, $\alpha_{ck}(\in (0,1])$, is the choice kernel learning rate. $C(t,j)=1$ if option j was selected at trial t, otherwise $C(t,j)=0$. Unlike expected value, choice kernel values for *all* options are updated on every trial.

$$
v_j(t+1) = v_j(t) + \delta_j(t) \cdot \alpha \cdot (O(t) - v_j(t)) \tag{5}
$$

$$
ck_j(t+1) = ck_j(t) + \alpha_{ck} \cdot (C(t,j) - ck_j(t)) \tag{6}
$$

The choice kernel modified softmax, Eq (7), extends the basic softmax of Eq (4) with an additional term in the exponents consisting of the product of a second inverse temperature, $\beta_{ck}(\in \mathbf{R}^+)$, and choice frequency estimates from Eq (6). Larger β_{ck} values in combination with higher choice frequencies will increase the probability that an option is chosen repeatedly. If β_{ck} is set to zero this rule reverts to the basic softmax Eq (4).

$$
P[C(t,i)=1] = \frac{e^{\beta \cdot v_i(t) + \beta_{ck} \cdot ck_i(t)}}{\sum_{j=1}^{N} e^{\beta \cdot v_j(t) + \beta_{ck} \cdot ck_j(t)}} \tag{7}
$$

Dual Learning Rate (DLR) RL

DLR RL was inspired by the "optimism bias", which suggests humans respond asymmetrically to rewards and punishment [24]. This model's learning rule comprises a set of equations, Eq (8). This learning rule is similar to the delta learning rule (Eq (5)) except the learning rate, $\alpha(\in [0,1])$, depends on the outcome of the current trial. If the outcome is a loss ($O(t)=0$), a learning rate associated with losses is used to update the value. If the outcome is a win ($O(t)=1$), a learning rate associated with wins is used to update the value. The DLR RL model uses the basic softmax, Eq (4), to model its decision process.

$$
v_j(t+1) = \begin{cases} v_j(t) + \delta_j(t) \cdot \alpha_{lss} \cdot (O(t) - v_j(t)) & \text{if O(t) = 0} \\ v_j(t) + \delta_j(t) \cdot \alpha_{win} \cdot (O(t) - v_j(t)) & \text{if O(t) = 1} \end{cases} \tag{8}
$$

Noisy Learning RL

The Noisy Learning RL model, derived by Findling and Colleagues [25], employs a learning rule, Eq (9), similar to the delta learning rule in Eq (5). However, its learning rule adds a time-varying noise term, $\varepsilon(t)$. $\varepsilon(t)$ is a zero-mean, normally distributed noise component, $\varepsilon(t) \sim \mathcal{N}(0, \sigma(t))$. The standard deviation, $\sigma(t)$, is equal to a constant fraction of the prediction error magnitude Eq (10). $\zeta(\in ??)$ is a free parameter of the model. When $\zeta = 0$, this rule is equal to the delta learning rule of Eq (5).

$$v_j(t+1) = v_j(t) + \delta_j(t) \cdot \alpha \cdot (O(t) - v_j(t)) + \varepsilon(t) \tag{9}$$

$$\sigma(t) = \zeta |O(t) - v_j(t)| \tag{10}$$

The Noisy RL model uses the decision rule presented in Eq 11. It is the basic softmax in Eq (11) with an added "choice hysteresis" term in the exponential. The choice hysteresis $\xi(\in ??)$ is a free parameter of the model and serves a similar purpose to the choice kernel in Eq (7) [1]; it increases the probability of an option being chosen during the current trial if it was chosen on the previous trial. Specifically, $\delta_j(t-1) = 1$ if option j was chosen on trial t-1 and is 0 otherwise and $\delta_i(t-1) = 1$ if option i was chosen on trial t-1 and is 0 otherwise. If ξ is set to zero this decision rule reverts to the basic softmax. Note the equation as presented here is slightly modified from how it is written in Findling et. al [25] by using a delta function for the choice hysteresis term due to our study using a 3-arm bandit versus a 2-arm bandit they used.

$$P(C(t,i) = 1) = \frac{e^{\beta \cdot v_i(t) + \xi \cdot \delta_i(t-1)}}{\sum_{j=1}^{N} e^{\beta \cdot v_j(t) + \xi \cdot \delta_j(t-1)}} \tag{11}$$

Exploration Bonus RL

The exploration bonus RL model was employed by Daw and Colleagues [26] in their 2006 investigation into human decision making. In their study, the model included a Kalman filter based learning rule, however we used the basic delta learning rule (eq 5) for consistency with other models in the comparison, as well as the fact decision processes are the main focus here. The exploration bonus model's decision process is a modified softmax, Eq (12).

$$P(C(t,i) = 1) = \frac{e^{\beta \cdot [v_i(t) + \varphi \cdot E_i(t)]}}{\sum_{j=1}^{N} e^{\beta \cdot [v_j(t) + \varphi \cdot E_j(t)]}} \tag{12}$$

Table 2. Summary and comparison of value-comparison RL models' parameters and equations

Model	Learning Rule	Policy	Free Parameters
Dual Learning Rate	dual delta	basic softmax	α_{win}, α_{loss}, β
Choice Kernel	value delta + ck delta	softmax w/ ck extended exponential	α, β, α_{ck}, β_{ck}
Noisy Learning	delta w noise component	softmax w/ choice hysteresis	α, β, ζ, ξ
Exploration Bonus	delta	softmax w/ eb value add	α, β, φ

[1] Note that if $\alpha_{ck} = 1$ in the choice kernel learning rule of Eq (6), the choice kernel of the currently selected option will immediately go to one, while the choice kernel's of the unselected options are immediately set to zero. Therefore, when $\alpha_{ck} = 1$ the choice kernel softmax and the choice hysteresis softmax rules are functionally equivalent (modulo β_{ck} and ξ).

Behavioral (Self Reported) Measures

4 participants out of 1000 were excluded from this analysis due to inconsistencies in the self-reported symptom measures. Each of the remaining participants' model parameters were extracted and z-scored prior to further analysis.

A Poisson generalized linear model (GLM) was fit to a variety of outcome measures separately for each model's parameter set. Specifically, for generalized anxiety symptoms and behavioral motivation **(Figure 4a)**, we first made a "best-parameter model" by fitting a univariate GLM to the one parameter that produced the highest R^2 (e.g. inverse temperature (β) for apathy motivation in the RLCK model) out of all parameters for both RLCK and foraging model. We then constructed a "total-parameter model" by including all parameters from the RLCK model (α, β, αck, βck) or from the foraging model (α, β, ρ) in a multivariate poisson GLM. We compared the R^2 of the best-parameter model to that of the total-parameter model to see how much extra variance the full set of parameters explained.

We next extended the total-parameter model framework to eleven additional clinical and cognitive symptom scores, where we computed the Akaike Information Criterion:

$$\mathrm{AIC} = 2\,k \;-\; 2\,\mathrm{LL} \tag{13}$$

where k is the number of estimated regression coefficients (including the intercept) and LL is the Poisson log-likelihood evaluated at the fitted parameters. This permitted us to measure the AIC differences for each model relative to the best fit:

$$\Delta_i \;=\; \mathrm{AIC}_i \;-\; \min_{j\in\{\mathrm{RLCK},\,\mathrm{FOR}\}} \mathrm{AIC}_j \tag{14}$$

Finally, using these differences in AIC, we were next able to convert these values into the Akaike weights, which serves to estimate the relative likelihood of each model. This let us compare weights for either model in **Figure 4b** :

$$w_i \;=\; \frac{\exp\!\left(-\frac{1}{2}\,\Delta_i\right)}{\displaystyle\sum_{j\in\{\mathrm{RLCK},\,\mathrm{FOR}\}} \exp\!\left(-\frac{1}{2}\,\Delta_j\right)}$$

References

1. Addicott MA, Pearson JM, Sweitzer MM, Barack DL, Platt ML. A Primer on Foraging and the Explore/Exploit Trade-Off for Psychiatry Research. Neuropsychopharmacol. 2017;42(10):1931–1939. doi:10.1038/npp.2017.108.

2. Kolling N, Akam T. (Reinforcement?) Learning to forage optimally. Current Opinion in Neurobiology. 2017;46:162–169. doi:10.1016/j.conb.2017.08.008.

3. Charnov EL. Optimal foraging, the marginal value theorem. Theoretical Population Biology. 1976;9(2):129–136. doi:10.1016/0040-5809(76)90040-X.

4. Constantino SM, Daw ND. Learning the opportunity cost of time in a patch-foraging task. Cogn Affect Behav Neurosci. 2015;15(4):837–853. doi:10.3758/s13415-015-0350-y.

5. Wilson RC, Collins AG. Ten simple rules for the computational modeling of behavioral data. eLife. 2019;8:e49547. doi:10.7554/eLife.49547.

6. Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. In: Parzen E, Tanabe K, Kitagawa G, editors. Selected Papers of Hirotugu Akaike. Springer Series in Statistics. New York, NY: Springer; 1998. p. 199–213. Available from: `https://doi.org/10.1007/978-1-4612-1694-0_15`.

7. Wagenmakers EJ, Farrell S. AIC model selection using Akaike weights. Psychonomic Bulletin & Review. 2004;11(1):192–196. doi:10.3758/BF03206482.

8. Palminteri S, Wyart V, Koechlin E. The Importance of Falsification in Computational Cognitive Modeling. Trends in Cognitive Sciences. 2017;21(6):425–433. doi:10.1016/j.tics.2017.03.011.

9. Wagenmakers EJ, Ratcliff R, Gomez P, Iverson GJ. Assessing model mimicry using the parametric bootstrap;48(1):28–50. doi:10.1016/j.jmp.2003.11.004.

10. Worthy DA, Hawthorne MJ, Otto AR. Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models. Psychon Bull Rev. 2013;20(2):364–371. doi:10.3758/s13423-012-0324-9.

11. Nowak M, Sigmund K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature. 1993;364(6432):56–58. doi:10.1038/364056a0.

12. Goodnow JJ, Pettigrew TF. Effect of prior patterns of experience upon strategies and learning sets. Journal of Experimental Psychology. 1955;49(6):381–389. doi:10.1037/h0049350.

13. Otto AR, Taylor EG, Markman AB. There are at least two kinds of probability matching: Evidence from a secondary task. Cognition. 2011;118(2):274–279. doi:10.1016/j.cognition.2010.11.009.

14. Steyvers M, Lee MD, Wagenmakers EJ. A Bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology. 2009;53(3):168–179. doi:10.1016/j.jmp.2008.11.002.

15. Paulus MP, Huys QJM, Maia TV. A Roadmap for the Development of Applied Computational Psychiatry. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. 2016;1(5):386–392. doi:https://doi.org/10.1016/j.bpsc.2016.05.001.

16. Wang XJ, Krystal JH. Computational Psychiatry. Neuron. 2014;84(3):638–654. doi:10.1016/j.neuron.2014.10.018.

17. Geana A, Wilson RC, Daw N, Cohen JD. Information-Seeking, Learning and the Marginal Value Theorem: A Normative Approach to Adaptive Exploration. Proceedings of the Annual Meeting of the Cognitive Science Society. 2016;38.

18. Lee D, Seo H, Jung MW. Neural Basis of Reinforcement Learning and Decision Making. Annu Rev Neurosci. 2012;35(1):287–308. doi:10.1146/annurev-neuro-062111-150512.

19. Eckstein MK, Wilbrecht L, Collins AG. What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. Current Opinion in Behavioral Sciences. 2021;41:128–137. doi:10.1016/j.cobeha.2021.06.004.

20. Fan H, Gershman SJ, Phelps EA. Trait somatic anxiety is associated with reduced directed exploration and underestimation of uncertainty. Nat Hum Behav. 2022;7(1):102–113. doi:10.1038/s41562-022-01455-y.

21. Scholl J, Trier HA, Rushworth MFS, Kolling N. The effect of apathy and compulsivity on planning and stopping in sequential decision-making. PLoS Biol. 2022;20(3):e3001566. doi:10.1371/journal.pbio.3001566.

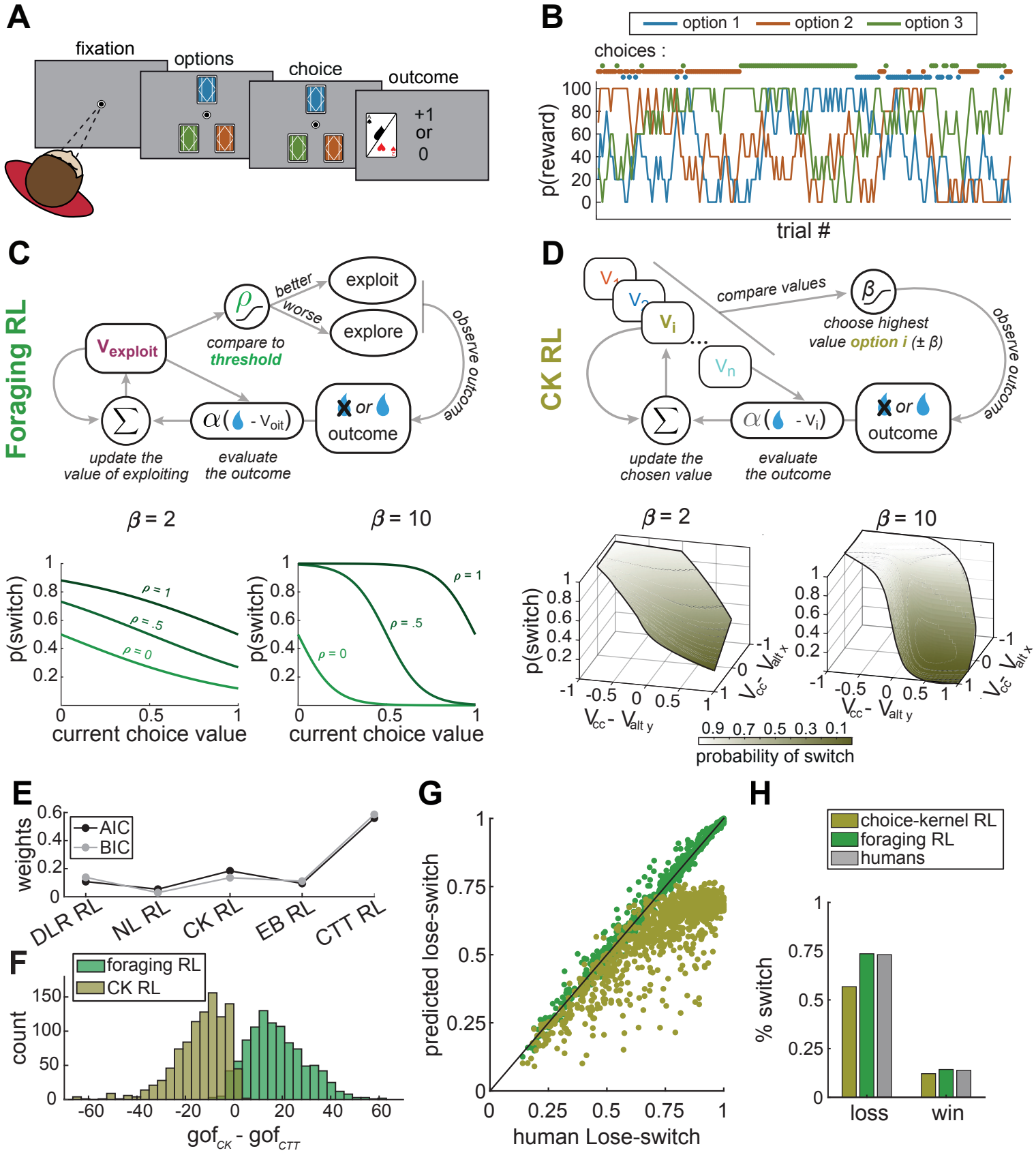
22. Sutton RS, Barto AG. Reinforcement learning: an introduction. Second edition ed. Adaptive computation and machine learning series. Cambridge, Massachusetts: The MIT Press; 2018.

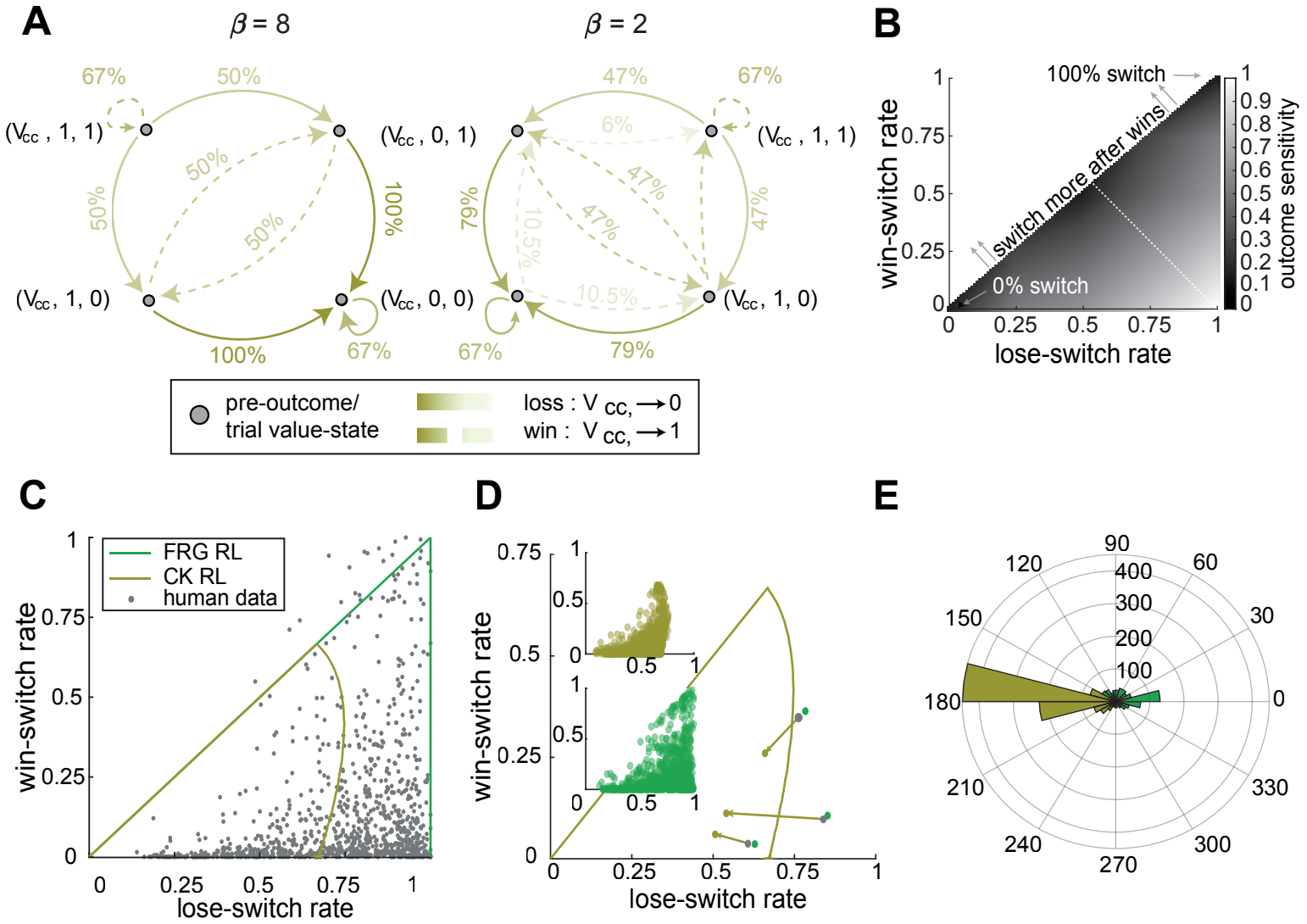
23. Zhang L, Lengersdorff L, Mikus N, Gläscher J, Lamm C. Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. Social Cognitive and Affective Neuroscience. 2020;15(6):695–707. doi:10.1093/scan/nsaa089.

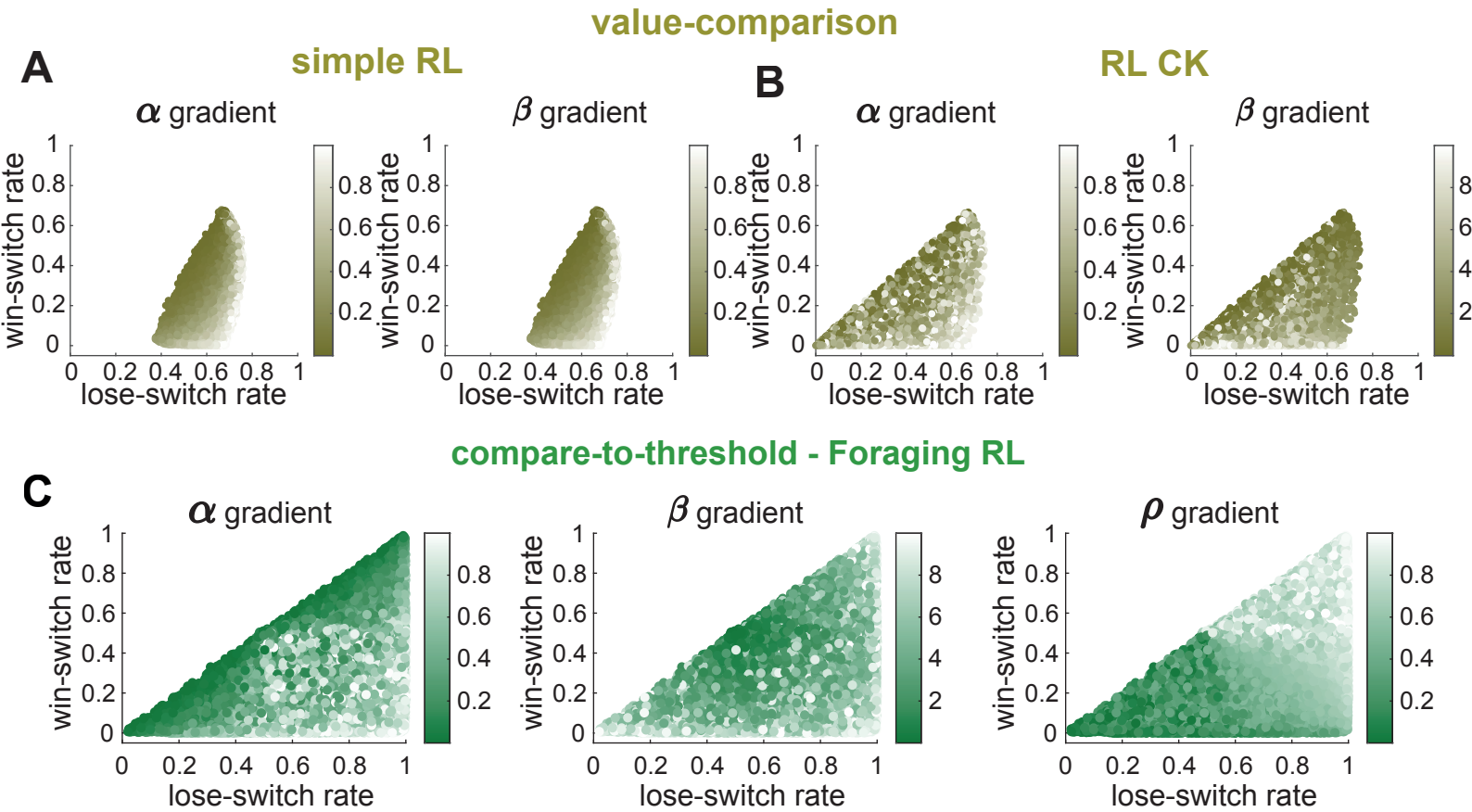
24. Kandroodi MR, Vahabie AH, Ahmadi S, Araabi BN, Ahmadabadi MN. Optimal Reinforcement Learning with Asymmetric Updating in Volatile Environments: a Simulation Study. Neuroscience; 2021. Available from: `http://biorxiv.org/lookup/doi/10.1101/2021.02.15.431283`.

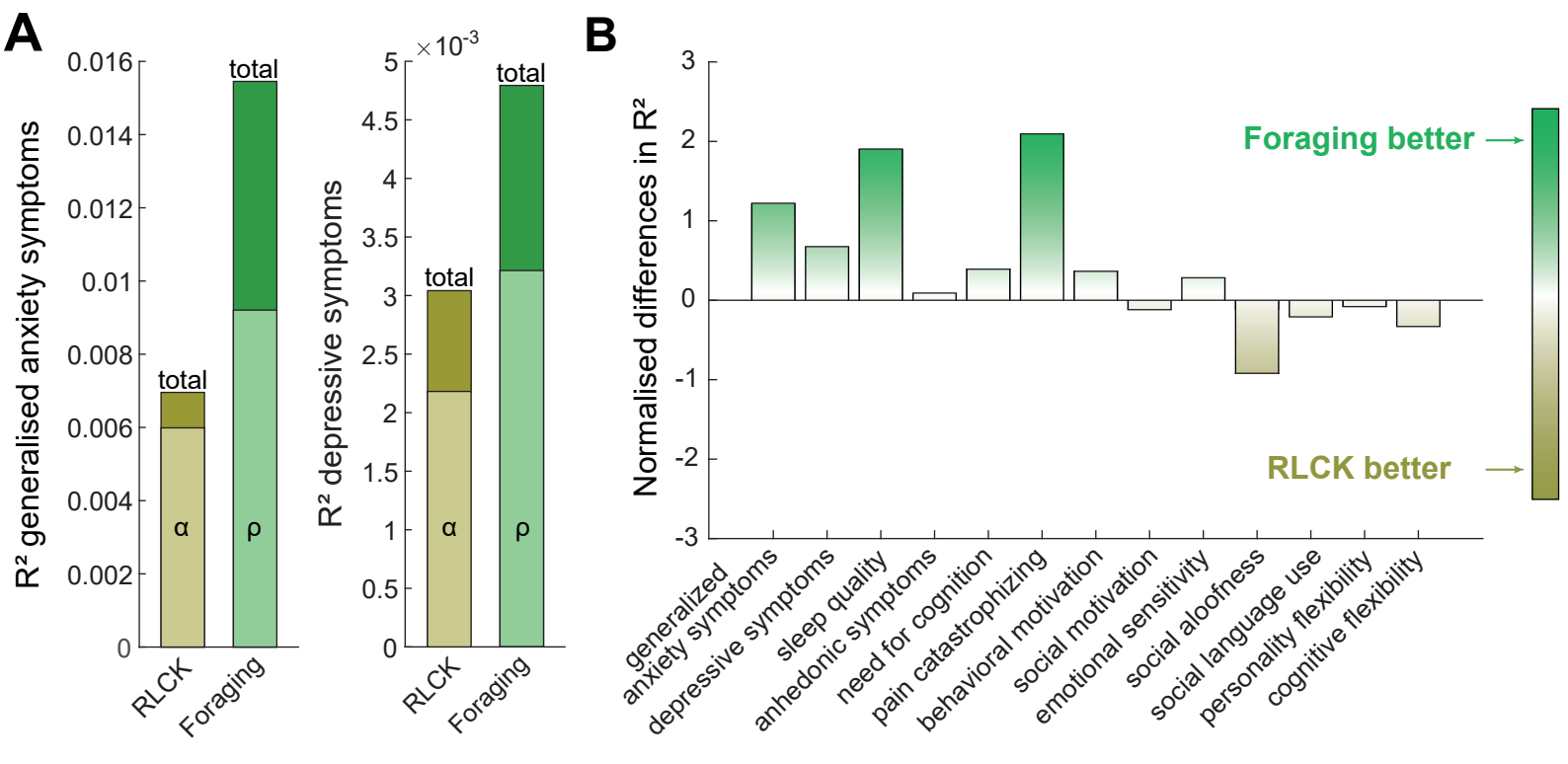
25. Findling C, Skvortsova V, Dromnelle R, Palminteri S, Wyart V. Computational noise in reward-guided learning drives behavioral variability in volatile environments. Nat Neurosci. 2019;22(12):2066–2077. doi:10.1038/s41593-019-0518-9.

26. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006;441(7095):876–879. doi:10.1038/nature04766.


A
fixation
options
choice
outcome
+1
or
0
B
option 1
option 2
option 3
choices :
p(reward)
trial #
C
Foraging RL
better
worse
exploit
explore
compare to
threshold
observe outcome
update the
value of exploiting
evaluate
the outcome
outcome
D
CK RL
compare values
choose highest
value option i (± β)
update the
chosen value
evaluate
the outcome
β = 2
β = 10
p(switch)
current choice value
probability of switch
E
weights
AIC
BIC
DLR RL
NL RL
CK RL
EB RL
CTT RL
F
foraging RL
CK RL
count
G
predicted lose-switch
human Lose-switch
H
choice-kernel RL
foraging RL
humans
% switch
loss
win


A
β = 8
β = 2
67%
50%
(Vcc , 1, 1)
(Vcc , 0, 1)
50%
50%
50%
100%
(Vcc , 1, 0)
(Vcc , 0, 0)
100%
67%
47%
67%
6%
(Vcc , 1, 1)
47%
47%
47%
79%
10.5%
10.5%
(Vcc , 1, 0)
67%
79%
pre-outcome/
trial value-state
loss : V cc, → 0
win : V cc, → 1
B
100% switch
switch more after wins
0% switch
win-switch rate
lose-switch rate
outcome sensitivity
C
FRG RL
CK RL
human data
win-switch rate
lose-switch rate
D
win-switch rate
lose-switch rate
E


value-comparison
A
simple RL
α gradient
β gradient
B
RL CK
α gradient
β gradient
win-switch rate
lose-switch rate
compare-to-threshold - Foraging RL
C
α gradient
β gradient
ρ gradient
win-switch rate
lose-switch rate


A
B
R² generalised anxiety symptoms
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
total
total
α
ρ
RLCK
Foraging
R² depressive symptoms
×10-3
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
total
total
α
ρ
RLCK
Foraging
Normalised differences in R²
3
2
1
0
-1
-2
-3
Foraging better
RLCK better
generalized anxiety symptoms
depressive symptoms
sleep quality
anhedonic symptoms
need for cognition
pain catastrophizing
behavioral motivation
social motivation
emotional sensitivity
social aloofness
social language use
personality flexibility
cognitive flexibility