Humans forage for reward in reinforcement learning tasks

Meriam Zid[1*], Veldon-James Laurie[1], Alix Levine-Champagne[1], Akram Shourkeshti[1], Dameon Harrell[2], Alexander B. Herman[2†], R. Becket Ebitz[1*†]

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

*Corresponding author(s). E-mail(s): meriam.zid@umontreal.ca; becket.ebitz@umontreal.ca;
†These authors contributed equally to this work.

Abstract

How do we make good decisions in uncertain environments? In psychology and neuroscience, the classic answer is that we calculate the value of each option and then compare the values to choose the most rewarding, modulo some exploratory noise. An ethologist, conversely, would argue that we commit to one option until its value drops below a threshold, at which point we start exploring other options. In order to determine which view better describes human decision-making, we developed a novel, foraging-inspired sequential decision-making model and used it to ask whether humans compare to threshold ("Forage") or compare alternatives ("Reinforcement-Learn" [RL]). We found that the foraging model was a better fit for participant behavior, better predicted the participants' tendency to repeat choices, and predicted the existence of held-out participants with a pattern of choice that was almost impossible under RL. Together, these results suggest that humans use foraging computations, rather than RL, even in classic reinforcement learning tasks.

Keywords: foraging theory, reinforcement learning, decision-making, human behavior

Introduction

Because the world is only imperfectly observable and frequently changes, many of the decisions we make are uncertain. How do we make good decisions despite this uncertainty? The conventional model in psychology and neuroscience is that we (1) track the subjective utility or "value" of each option within a set and then (2) compare option values to make the best possible decision [1–5]. The "compare-alternatives" perspective lies at the heart of the **Reinforcement Learning** (RL) models that are ubiquitous in the cognitive sciences [1, 6]. It also fundamentally shapes the way that tasks are designed in these fields. Because comparing alternatives requires that all values are known at the time of the choice, for example, laboratory task typically present all options simultaneously [7–10]. Similarly, because the reward prediction error in RL models functions as to integrate rewards over time in order to track changing values [1, 6], option values often have complex, unpredictable dynamics that must be learned in psychology and neuroscience tasks.

While the compare-alternatives view is widely influential, it is not universally accepted. In ethology, for example, **foraging** models instead assume that we (1) calculate the value of one exploited option and (2) compare its value against a threshold to decide whether to continue exploiting or try something new [11, 12]. Although the foraging model may seem simplistic compared to RL, the threshold against which exploited options are compared can integrate factors like the value of alternative options (or the "richness" of the environment) but also travel, time and handling costs associated with trying something new [13]. The foraging perspective has begun to gain traction among human cognitive scientists, but because of certain limitations in how these models are formalized, foraging models are generally only fit in circumstances where reward values decay at a predictable rate and options are encountered in sequence—in tasks that mirror the assumptions at the heart of this model [14–19]. In short, because the compare-alternatives and compare-to-threshold views of sequential decision-making use incomparable computational models to make sense of incompatible behaviors, we do not know which view best describes human decision-making in standard laboratory tasks.

Here, we tested the idea that humans forage even in classic reinforcement learning tasks. We developed a novel sequential decision-making model that is inspired by foraging theory rather than the compare-alternatives approach at the heart of RL. Across 3 independent experiments, we compared this new foraging model with various variations on RL models [4]. We found that the foraging model was a better fit for the participants' decisions, outperformed multiple variations on the RL model, and better predicted the participants' tendency to repeat choices on both individual and group level. Together, these results suggest that humans tend to use foraging-like algorithms, over RL-like algorithms, even in canonical RL tasks. In short, humans compare to threshold rather than compare alternatives.

Results

Participants performed above chance and used complex strategies

In Experiment 1, participants on the Amazon mTurk platform (258 people, 120 female, 2 other or non-reporting) performed a classic sequential decision-making task known as a restless k-armed bandit (Figure 1A; [6, 20–24]). The participants were asked to repeatedly choose between k options, each of which was associated with some probability of reward. Reward probabilities changed unpredictably over time and independently across options (Figure 1B). The reward structure was not cued to the participants so the only way to infer the value of an option was to sample it and find out. Because rewards evolved over time, the longer the participants went without sampling an option, the more uncertain its payoff became. This task naturally encourages decision-makers to alternate between exploiting valuable options and exploring uncertain alternatives because the latter could become better at any time. It is a classic in the RL literature and has become a testbed for evaluating RL models in both psychology and artificial intelligence [20, 25, 26].

Participants were generally good at the task, despite its uncertainty. They chose the objectively best option 76.6% of the time (+/- 11.5% STD; Figure 1C) and received rewards 19.2 % more frequently than would be expected by chance (+/- 15.3 % STD; Figure 1D). (Note that 4/258 participants [1.6%] were excluded from this and further analyses because they chose only one option; but see Figure 1C-E for analyses of these 4 participants.) Participants were more likely to repeat choices than to switch (switching on only 19.9% of trials, +/- 14.5% STD). They were also more likely to repeat after reward (win-stay = 93.3%, +/- 11.1% STD) and switch after no-reward (lose-shift = 39.2%, +/- 21.0% STD). However, no participants followed a deterministic win-stay-lose-shift rule (all participants either win-shifted or lose-stayed at least 5% of the time).

Foraging models better explain participant decisions

To determine how the participants were solving this task, we fit computational models (see Methods). We first compared a simple 2-parameter Reinforcement learning model ("standard RL"; Rescorla-Wagner; [1, 6]) with a simple 3-parameter formulation of the foraging model ("standard F"). Although both models track the local rate of reward via identical delta-rule computations, the foraging model decides whether to stay or switch via comparing the value of the exploited option against a threshold (compare-to-threshold; Figure 2A), while the RL model decides between options via comparing values across the option set (value-comparison; Figure 2B). We found that the standard foraging model was a better fit to participant behavior than the standard RL model (Figure 2C-D; standard F: log-likelihood = -23,206, AIC = 47,936, 3 parameters by 254 participants = 762 parameters; standard RL: log-likelihood = -24,878, AIC = 50,771, relative AIC weight $< 1.8 * 10^{-32}$, 2 x 254 participants = 508 parameters; 76,200 total trials). The foraging model also outperformed the RL model on an individual basis: it was a better fit in 64.2% of individual participants (163/254; Figure 2E; standard

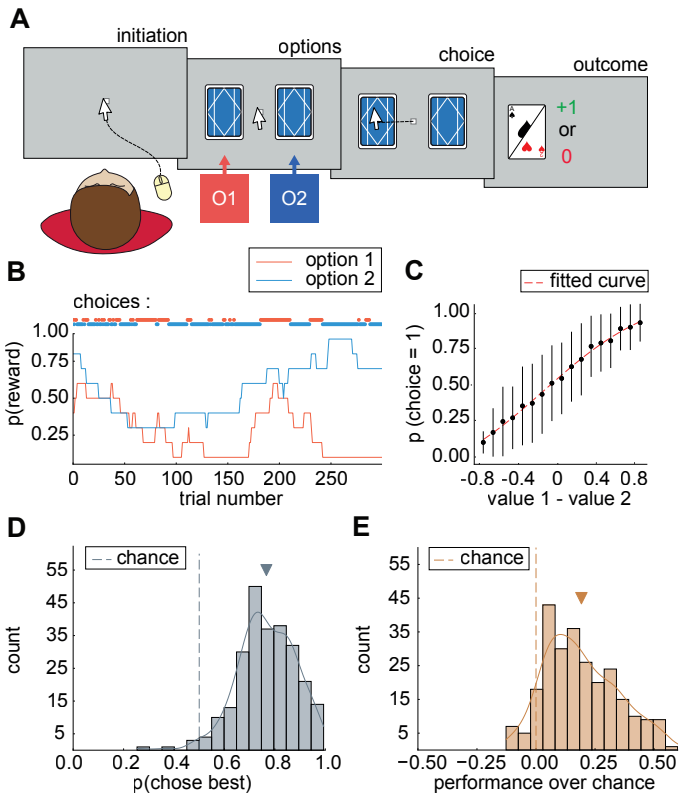


Fig. 1: Task and baseline behavior. A) Participants (n=258) chose between 2 face-down decks of playing cards that were composed of aces and 2's. Participants were told that the proportion of aces and 2's in each deck would change over time and that decks could be good, bad, or mediocre right now, but would not stay that way the whole time. Participants got a point (and $0.02) for each ace they found, but no points for 2's. **B)** The reward probability of each option (red trace = option 1; blue trace = option from an example reward schedule, with participant choices along the top (red dots = option 1; blue dots = option 2). **C)** Average probability of choosing option 1 as a function of the difference in the objective reward probability ("value") between option 1 and 2. Error bars = standard error of the mean (SEM) across participants. **D)** Distribution of the number of trials in which each participant chose the objectively best option. Dotted line = chance, carrot = mean across participants.**E)** Same as *D*, for the percent of trials in which participants were rewarded, normalized such that units represent percent of chance performance.

F: median individual log-likelihood = 82.36, 95%CI = 9.40 to 201.06, standard RL: median = 86.72, 95% CI = 9.56 to 199.66; significant paired t-test: $p < 0.0001$, t(253) = 6.79, mean difference = 5.27, 95% CI = 3.74 to 6.79). This suggests that a foraging-like compare-to-threshold mechanism may better describe the participants' choices in this task.

Next, we asked whether the foraging model also outperforms a variety of extended RL models previously shown to outperform the standard 2-parameter formulation (Figure 2F-G). Every extended RL model performed substantially better than the

standard RL model in model comparison. However, the foraging model continued to outperform: it was a better fit than RL models that incorporated asymmetrical learning from wins and losses ("asymmetrical RL" [23, 27, 28]; log-likelihood = -23,963, AIC = 49,450, AIC weight relative to foraging model $< 10^{-32}$, 762 parameters), decay in the value of unchosen options ("decaying RL"; [21, 29]; log-likelihood = -23,181, AIC = 47,887, relative AIC weight $< 10^{-32}$, 762 parameters), and a simple form of choice-history dependence with 1 added parameter ("history-kernel 1 RL"; [4]; log-likelihood = -23,252, AIC = 48,028, AIC weight relative to foraging model $< 10^{-32}$, 762 parameters). The only extended RL model that outperformed the simple foraging model was a 2-parameter choice-history dependent model ("history-kernel 2 RL"; [4]; log-likelihood = -22,660, AIC = 47,351, relative weight of the foraging model $< 10^{-32}$). Thus, the simplest foraging model was a better explanation for behavior than all but the most complex of RL models.

Although the best extended RL model outperformed the simplest foraging model, it was not clear if this was driven by the fact that it was a value-comparison model or by the fact that it was a history dependent model. If history dependence mattered, then adding history dependence to the foraging model should allow it to outperform even this best RL model. Indeed, the best history-dependent foraging model outperformed the best history-dependent RL model (Figure 2F-G; see Methods; "history-kernel 2 foraging": log-likelihood = -22,292, AIC = 47,125, AIC weight relative to choice-kernel RL 2 $< 10^{-32}$, 1270 parameters). Critically, this meant that each extended foraging model outperformed the equivalent RL model. This was true for a foraging model that permitted asymmetrical learning between wins and losses ("asymmetrical foraging": log-likelihood = -22,777, AIC = 47,585, AIC weight relative to asymmetrical RL $< 10^{-32}$, 4 x 254 = 1016 parameters), for a foraging model where the threshold decayed with repeated choices ("decaying foraging": log-likelihood = -22,632, AIC = 47,295, AIC weight relative to decaying RL $< 10^{-32}$, 1016 parameters), and for the 1-parameter history-dependent model ("history-kernel 1 foraging": log-likelihood = -22,558, AIC = 47, 148, AIC weight relative to choice-kernel RL 1 $< 10^{-32}$, 1016 parameters). In short, the best foraging model was the best overall and, as a group, the foraging models consistently provided a better fit to the participants' decisions.

The overall best fitting model was a 5-parameter foraging model that included 2 choice-history dependent parameters (relative AIC weights of all competing models $< 10^{-32}$). This foraging model also outperformed the best RL model on an individual basis: it did at least as well or better in every individual participant than the best RL model (foraging model: median individual log-likelihood = 79.32, 95% CI = 8.92 to 197.18, RL model: median = 80.43, 95% CI = 8.72 to 195.31; significant paired t-test: $p < 0.05$, t(253) = 1.98, mean difference = 1.45, 95% CI = 0.01 to 2.89). Together, these results suggest that foraging theory's compare-to-threshold decision-making better described human choice behavior than value-comparison did.

RL outperformed foraging on the task

Although the k-armed bandit is a classic testbed in the RL literature [4, 6, 20, 30], it was nonetheless possible that some aspect of the task design encouraged the participants to use foraging rather than RL computations. To determine if this was the

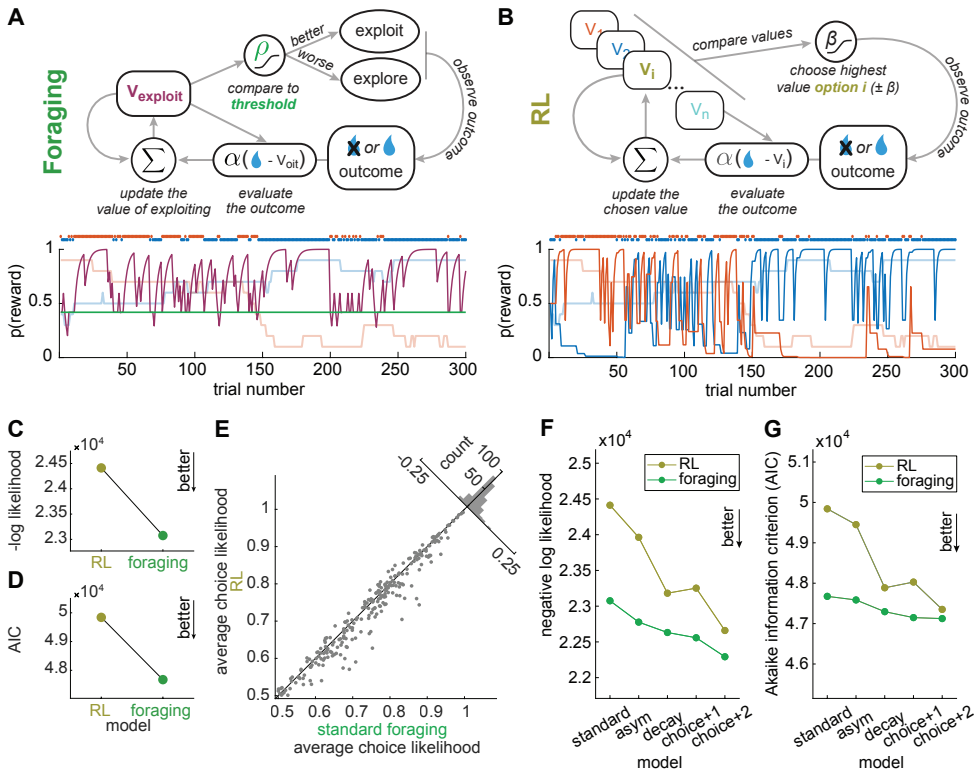


Fig. 2: Foraging and RL models produce different choice patterns and fits to the data. A) *Top*: A cartoon illustrating the mechanics of the foraging model. The model estimates the value of exploiting a current favored option through a standard delta-rule update. The value of exploitation is then compared to a threshold in order to decide whether to continue exploiting or to begin to explore. *Bottom*: Simulated data from the foraging model, fit to the example participant and run on the example reward schedule. The choices (colored dots) and exploitation value (purple trace) of the foraging model are shown, along with the objective reward probabilities (pale traces). The horizontal green line is the threshold. **B)** Same as A for the RL model. Here, each option is assigned its own value and values are tracked independently. Option values are then compared in order to choose what to do in the next trial. *Bottom*: Simulated data from the RL model. The darker blue and red colored traces represent the subjective value of the two options during the simulation. **C)** The negative log-likelihood of the RL and foraging models fit to all data from Experiment 1. **D)** The Akaike information criteria (AIC) of the RL and foraging models. **E)** Average choice likelihood for each participant under the foraging model (x-axis) and the RL model (y-axis). *Inset*: Distribution of differences in choice likelihood across participants. **F)** The negative log-likelihood of various extensions of the RL (gold) and foraging models (green). **G)** The Akaike information criteria of the extensions of the RL (gold) and foraging models (green).

case, we compared the performance of the optimal foraging and RL agents in the task (Figure 3A-D; see Methods). We found that the environment did not advantage foraging over RL; in fact, RL agents outperformed foraging agents, regardless of whether performance was defined as the probability of reward (wilcoxon signed rank test : $p < 0.01$; Figure 3 C/D) or the probability of choosing the objectively best option (wilcoxon signed rank test : $p < 0.01$; Figure 3A/B). There was no relationship between how foraging-like an individual participant was and how well they did in the task. Individual differences in the "foraging index" (see Methods) were

not correlated with the probability that the participant would choose the best option ($R = -0.01, p = 0.84$) and they were not correlated with above-chance reward probability ($R = 0.03, p = 0.64$). In sum, using foraging computations did not create an advantage or a disadvantage in this task, either in theory or in practice. It is thus not likely that the participants used a foraging strategy because the task somehow encouraged compare-to-threshold decision-making.

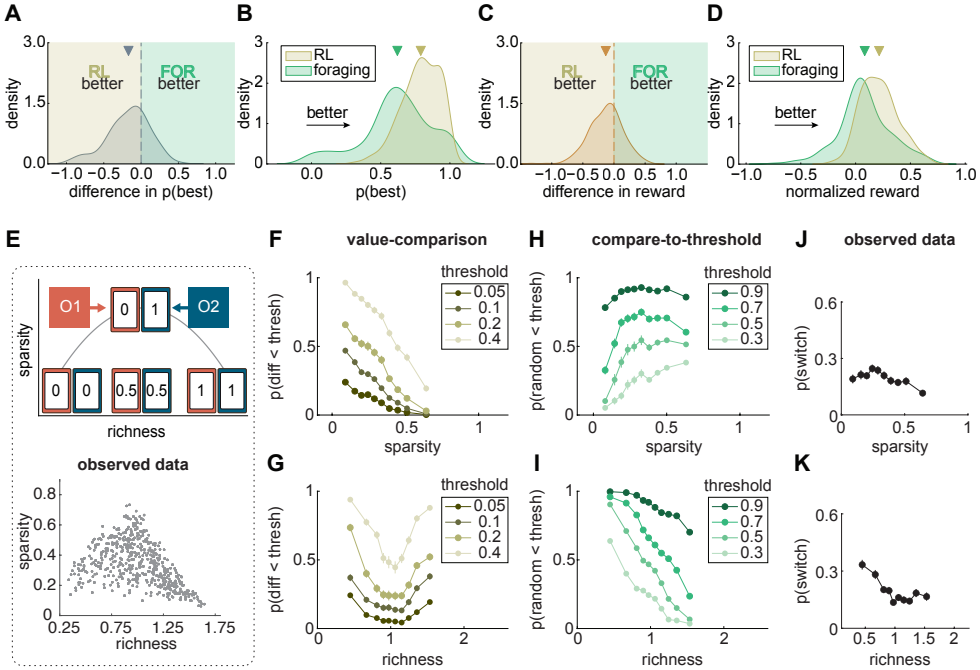


Fig. 3: Compare-to-threshold and value-comparison agents have differences in performance and dissociable signatures in rich and sparse environments. A) Distribution of the probability of choosing the objectively best option for both foraging (green) and RL (gold) agents. **B)** Difference between the distributions in *A* for agents interacting with matched environments. **C-D)** Same as in *A-B* for the probability of reward. **E)** *Top*: Cartoon illustrating the non-linear relationship between sparsity (the difference between option values) and richness (sum of option values) in a bounded environment. *Bottom*: Sparsity as a function of richness in Experiment 1. Each dot represents a segment of 100 trials experienced by a participant. **F-G)** In value-comparison, switching behavior is maximal when the absolute difference in rewards is low [23]. If we take the difference in rewards as a proxy for the probability of switch, then the value-comparison hypothesis makes different predictions for how switching should change as a function of sparsity (*F*) and richness (*G*). Shades of gold illustrate different hypothetical "thresholds" for switching. **H-I)** In compare-to-threshold, switching behavior is maximal when the sum of rewards is low. Here, each option's value is individually compared to a preset threshold. If the average of these below-threshold comparison serves as a proxy for switching then the compare-to-threshold hypothesis makes different predictions for how switching should change as a function of sparsity (*H*) and richness (*I*). Shades of green illustrate different hypothetical "thresholds" for switching. **J-K)** Probability of switch of all participants in Experiment 1 as a function of different levels of sparsity *(J)* and richness *(K)* of the environment (errorbar = SEM).

Humans resemble the fingerprint of compare-to-threshold decisions, not value-comparison

Although this fact is only explicit in the foraging model, both foraging and RL balance two distinct goals: they exploit high-value options when those are available, but also occasionally switch to alternatives that could be better. Where the models differ is in how they decide to switch options (Figure 2A-B). In RL models, switching is generated by noise and happens most when it is difficult to discern which option is the best. This means that it should occur most frequently when the values of the arms are close together (i.e. in dense, rather than sparse environments; (Figure 2B). In foraging, conversely, switching happens when an exploited arm falls below a threshold. This means that it should occur most frequently when the value of any arm is most likely to be below the threshold (i.e. in poor, rather than rich environments; Figure 2A). In short, switching should depend on very different aspects of the environment in compare-to-threshold decision-making and value-comparison decision-making.

If richness and sparsity were orthogonal, richness would not affect switching in value-comparison decisions and sparsity would not affect switching in compare-to-threshold decisions. However, in this task, sparsity and richness had a complex, nonlinear relationship with each other (Figure 3E). Because reward probabilities are bounded, the environment can only be very poor (sum close to 0) or very rich (sum close to the number of options) when the option values are not sparse (i.e. difficult to discriminate because they are all either low or high). Because sparsity makes switching less likely in value-comparison decisions (Figure 3F), the fact that sparsity is maximal at intermediate richness implies that switching should be least likely at intermediate richness. That is, switching should have an inverted u-shaped relationship with richness in this task during value-comparison decision-making (Figure 3G). Similarly, because richness makes switching less likely in compare-to-threshold decisions (Figure 3I), the relationship between richness and sparsity implies that switching should have a complex and occasionally inverted u-shaped relationship with sparsity during compare-to-threshold decision-making (Figure 3H). Thus, although compare-to-threshold decisions are not based on sparsity and value-comparison decisions are not based on richness, each strategy should have a unique fingerprint as a function of these two environmental variables(Figure 3G-H).

Finally, we asked whether humans resembled the value-comparison or compare-to-threshold fingerprints. We found the compare-to-threshold fingerprint in human behavior (Figure 3J-K). The participants switched options the most in poor environments and at intermediate levels of sparsity, exactly as we would expect from a compare-to-threshold process, but not a value-comparison one.

Foraging better explained the participants' tendency to repeat

Because the RL model also produced compare-to-threshold behavior, we had to continue searching for the reason why the foraging model was a better fit to the participants. In a different primate species, we previously found that choices are considerably more repetitive than what can be explained with standard RL models [21, 31]. Further, each extension of the RL models that improved model fit had the effect of making

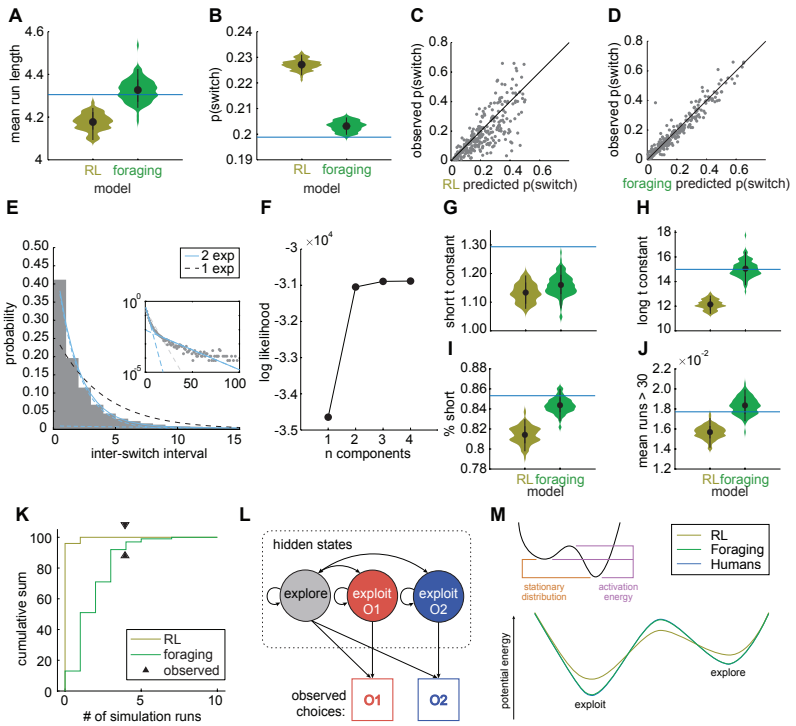


Fig. 4: Foraging and RL predict different choice dynamics. A) Average number of trials between switch decisions (run length) of each agent in simulated bandit environment and humans in Experiment 1 (blue line). **B)** Probability of switch. Same convention as in *A*. **C)** Human observed probability of switch as a function of RL predicted probability of switch. **D)** Foraging predicted probability of switch. Same convention as in *C*. **E)** Distribution of choice run lengths of humans in Experiment 1. If the probability of switching was fixed, run lengths would be exponentially distributed (black dotted line). A mixture of two exponential distributions (blue line) suggests 2 distinct probabilities of switching. Dotted blue lines show each mixing distribution, one slow-switching and another fast-switching. *Insert*: Same data presented on a logarithmic scale. **F)** Log-likelihoods for different mixture models containing a range of 1 to 4 exponential distributions. **G)** Probability of switch in the short time constant run lengths. **H)** Probability of switch in the long time constant run lengths. **I)** Proportion of run lengths with short time constant. **J)** Proportion of run lengths exceeding 30 consecutive trials. **K)** Four participants were held out of model fits because they chose the same option 300 trials in a row. To determine the expected number of no-switch participants under each model, 100 datasets were generated from the distribution of fitted parameters (each with 254 simulated participants). Here, we plot the cumulative sum of different numbers of no-switch participants from these 100 simulations. The black carrot is the proportion of no-switch participants in the original dataset. **L)** Hidden Markov model (HMM) was used to infer the underlying internal state of the participants on each trial from their sequence of choices. The model included 3 hidden states, 2 exploitative states corresponding to each option and an exploratory state which participants could choose any of the options. **M)** Overlaid state dynamic landscapes for both foraging (green), RL (gold) agents and humans (blue).

the model more repetitive. This is most obvious in the case of the choice history kernel models: the added parameters increase the probability of repeating the previous choice. In the decay model, similarly, repetition increases as the value of the alternative decays towards zero. Even in the asymmetrical learning models, learning more

from wins than losses tends to stabilize choices and increase repetition [32, 33]. Therefore, we next asked if the foraging model was a better fit to behavior because it better captured the repetitiveness of the participants' choices.

We generated simulated datasets that matched the participants from the standard RL and foraging models and then compared their level of repetitiveness with the participants (see Methods). In the participants, the average run length (length of same-choice sequences: 4.31 trials) was close to our expectation under foraging (4.33 trials, 95% CI = [4.24, 4.43]) but greater than 100% of the samples from the RL model (4.18 trials, 95% CI = [4.09, 4.24]; $p < 0.01$; Figure 4A). This was related to the fact that the foraging model was more accurate in predicting how often the participants would switch (i.e., the inverse of the choice run lengths; Figure 4B-D; foraging, mean sum of squared errors [MSSE] = 0.72, 95% CI across simulations = [0.53, 0.94], RL, MSSE = 2.78 (95% CI = [2.54,3.01]; significantly greater in RL; 2-sided paired t-test: $p < 0.0001$, t(1,253) = 6.46). The average probability of switching under the RL model was 22.7% (95% CI = [22.3%, 23.1%]) and under the foraging model it was 20.3% (95% CI = [20.0%, 20.7%]). The participants switched 20.0% of the time, less frequently than 100% of the samples from the RL model, but not the foraging model. In fact, the RL model systematically predicted that individual participants should switch more than they did (2.8% more, 95% CI across individuals = [1.7%, 4.0%], $p < 0.0001$, t(1,253) = 4.72). The foraging model did not have a significant bias towards over- or under-estimating individual participants (predicted 0.4% more on average, 95% CI = [-0.08% to 0.1%], p = 0.1, t(253) = 1.66). In short, in comparison to the RL model, the foraging model better predicted the participants' tendency to either repeat or switch on both the individual and group levels.

Foraging better explained persistent choice runs

In mice, monkeys, humans, and optimized RL models, the distribution of choice run lengths in this task is composed of two modes: one exploratory mode with rapid switching due to trial-and-error sampling and one exploitative mode involving persistent choices to the same target [21, 23, 31]. A specific increase in the persistent mode is what sets humans and other primates apart from rodents in this task [31] and it is also a feature of human decision-making that RL models do not naturally capture [21]. Therefore, we next asked whether the foraging model might better account for the persistent mode.

Again in these participants, run length distributions were best described as a mixture of 2 regimes (Figure 4E-F; see Exponential Mixture Models in Methods). The 2 regime model (log-likelihood = -31,048, 3 parameters, n = 14,847) provided a significantly better fit than 1 (log-likelihood = -34,645, 1 parameter, likelihood ratio test: ratio = 7193.0, p < 10^{-32}) and, while adding additional regimes continued to improve model fit (3 regimes: log-likehood = -30,893, 5 parameters; 4 regimes: log-likelihood = -30,884), improvement was already saturated at 2 (Figure 4F).

Two regimes were also apparent across all the simulated data from the RL and foraging models (RL: n = 1,699,789, 1 regime log-likelihood = -3,907,700, 2 regimes log likelihood = -3,527,300, 3 regimes log-likelihood = -3,502,000, 4 regimes log-likelihood = -3,500,600, significant improvement from 1 to 2: ratio = 760,950, p < 10^{-32}, elbow

at 2; foraging: n = 1,517,656, 1 regime log-likelihood = -3,550,000, 2 regimes log likelihood = -3,128,100, 3 regimes log-likelihood = -3,105,700, 4 regimes log-likelihood = -3,103,900, significant improvement from 1 to 2: ratio = 843,650, p $< 10^{-32}$, elbow at 2).

Although all three datasets had 2 switching regimes, those in the foraging model better matched the participants. Both the foraging and the RL model tended to switch more frequently than the humans did during their own exploratory regime (Figure 4G; RL = 1.13 trials, 95% CI = [1.08, 1.19]; foraging = 1.16, 95% CI = [1.08,1.23]; participants = 1.29, significantly longer than both models, $p < 0.01$). This may be due to the fact that neither model actually accounted for the complexity of exploratory decision-making, which can be information maximizing rather than random in some circumstances [34–36]. Nonetheless, only the foraging model was able to replicate the long switching constant of the participants' exploitative runs (Figure 4H; RL = 12.15 trials, 95% CI = [11.38, 12.86]; foraging = 15.05, 95% CI = [13.70,16.42]; participants = 14.99, only significantly different than RL [$p < 0.01$]) and their relative frequency (Figure 4J). This might explain why participants had more very long choice runs than we would expect under RL (1.8% of their choice runs were longer than 10% of the total number of trials [30 trials], significantly more than the 1.6% predicted by RL, 95% CI = [1.4%, 1.7%], $p < 0.01$) (Figure 4J). The participants' long-run frequency was well within the distribution under the foraging model, however (1.8%, 95% = [1.7%, 2.0%]). In sum, the foraging model better accounted for the participants tendency to exploit certain options for long periods of time.

Runs of choices that lasted the entire 300-trial duration of the session were very rare in RL simulations (4 of 25,400 or 0.016% of simulated sessions), in contrast to foraging (178 of 25,400 or 0.7% of simulated sessions). Recall that we initially excluded 4 of our 258 participants (1.6%) because they did exactly this, despite passing initial screening criteria. These represented held-out participants who did not influence the simulations in any way. Nonetheless, their choice sequences were much more likely under the foraging model than the RL model (Figure 4K). In fact, 4 or more participants like this would be expected in 10.9% of experiments under foraging. Under RL, we would have less than a 1 in 1 million chance of observing these 4 participants. In sum, the foraging model better captured the stability or repetitiveness of participants choice runs, in both those that were included in the model fits and in those that were excluded.

Foraging better explains the dynamics of exploration and exploitation

Foraging was better able to explain the participants' tendency to generate very long choice runs. This could suggest that the foraging model was simply more stable than the basic RL model. However, the foraging model also predicted that participants should have more fast-switching choices than the RL model did (Figure 4B/I; relative frequency of short choice runs; foraging = 84.4%, 95% CI = [82.8%, 86.0%]; RL = 81.4%, 95% CI = [79.8%, 83.0%]). Again the participants matched the predictions of the foraging model, but not the RL model: they had many more short choice runs than the RL model, but not the foraging model predicted (relative frequency of exploitative runs = 85.3%, only significantly different than RL [$p < 0.01$]). In short, the participants

both switched more and switched less than the RL model suggested they should, but foraging was able to capture these complex dynamics.

To understand why the foraging model was better at reproducing the participants' switching dynamics, we used latent state models to interrogate the dynamics of exploration and exploitation in the participants and in both models (see Methods; [21, 23, 35, 36]. Briefly, this approach models exploration and exploitation as the latent states underlying sequences of decisions through training hidden Markov models (HMMs) on observed choice sequences (Figure 4L). The models can be used to infer the most likely generative state underlying each individual choice [21, 23, 35], but they can also be used to make statistical inferences about the governing equations of exploration and exploitation [23, 31, 35]. For example, certain parameters in the transition matrix of an HMM are estimates of the long-term probability that a decision-maker will continue to exploit once they have begun to exploit, for example, or that they will stop exploiting one option in order to explore the alternatives.

The foraging model better matched the participants' explore/exploit state dynamics than did the RL model. Every measurement made with the HMMs was out of sample for the RL model, but well within sample of the foraging model. For example, the participants repeated exploration 90.4% of the time, which was significantly higher than the RL model simulations (87.3%, 95% CI = [86.3%, 88.4%], $p < 0.01$) but close to the mean of the foraging simulations (90.3%, 95% CI = [89.8%, 90.9%]). Similarly, the participants repeated exploitation 93.9% of the time, which was significantly higher than RL (90.9%, 95% CI = [90.5%, 91.3%], $p < 0.01$) but close to the mean foraging model (93.7%, 95% CI = [93.4%, 94.1%]). The fact that the transition matrix of an HMMs is a Markov chain makes it particularly analytically tractable. We can often solve for the stationary distributions of these equations: the long-term probability that the participants would exploit (vs. explore; see Methods). Here again, we found that the participants stationary frequency of exploration and exploitation (39.2% explore, 60.8% exploit) was out of sample for the RL model (41.6% explore, 95% CI = [40.4%, 42.9%]; 58.4% exploit, 95% CI = [57.1%, 59.6%], $p < 0.01$ in both cases), but well within sample for the foraging model (39.3% explore, 95% CI = [38.4%, 40.2%]; 60.7% exploit, 95% CI = [59.9%, 61.6%]). Together, these results illustrate that the foraging model well described the dynamics of exploration and exploitation in the participants, while the RL model was consistently off target.

Plotting the energetic landscape of these dynamics (see Methods; [23, 31, 35]) revealed the intuition behind these results: the energetic landscape of exploration and exploitation was flatter in the RL model than it was in either the participants or the foraging model (Figure 4M;). In both the participants and the foraging model, exploitation was a deeper, more stable state than exploration, but a significant energetic barrier between the two states meant that exploration and exploitation were actually fairly stable states, with infrequent transitions between them. Conversely, in the RL model, not only was there less of a difference in the depth of exploration and exploitation, but there was less of an energetic barrier between. Together, these results suggest that the foraging model was a better fit for the participants because it better captured the dynamics of exploration and exploitation: their tendency to alternate between temporally-extended periods of exploiting good options and exploring alternatives.

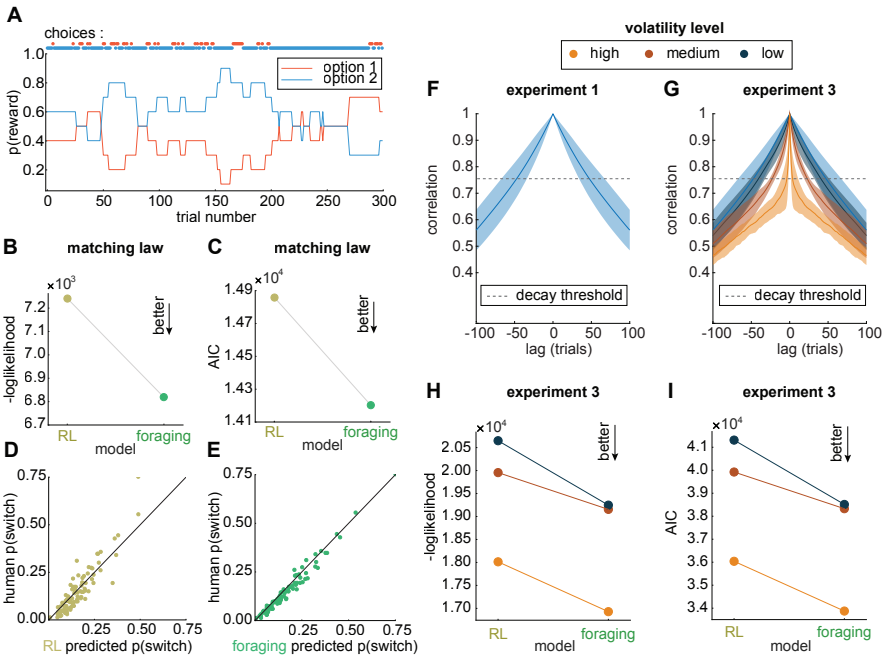


Fig. 5: Foraging and RL have different quality fit to the data across different environments. A) An example reward schedule from the matching law task (Experiment 2) , illustrating the reward probability of each option (red trace = option 1; blue trace = option 2) and the choices made by the participant who had this reward schedule (red dots = option 1; blue dots = option 2). **B)** The negative log-likelihood of the RL and foraging models fit to all data of Experiment 2. **C)** Akaike information criterion. Same convention as in *B*. **D)** Human observed probability of switch in Experiment 2 as a function of RL predicted probability of switch. **E)** Foraging predicted probability of switch. Same convention as in *D*. **F)** Reward schedules autocorrelations averaged for all participants of experiment 1. Full line indicates the mean, shading indicates STD and dotted line indicate the decay threshold. **G)** Same convention as in *F* for 3 levels of volatilities (red line) participants of experiment 3 **H)** The negative log-likelihood of the RL and foraging models fit to all data of Experiment 3. **I)** AIC of the RL and foraging models across different volatility levels in Experiment 3. Same convention as in *G*.

Foraging better predicts behavior under varied uncertainty

It is possible that the participants did not bother comparing alternative values because the value of unchosen options was difficult to estimate in a restless bandit environment. Therefore, in Experiment 2 and 3, we bidirectionally manipulated the uncertainty of the unchosen option.

In Experiment 2, participants (95 people, 41 females, 1 other or non reporting) performed an "matching law" version of the 2-armed bandit task. In this variation, the environment was structured so that the sum of reward probabilities for any given trial always totaled one ('Example walk', Figure 5A). This design ensured reward probabilities for the options fluctuated inversely and symmetrically, meaning that one choice was almost always more advantageous than the other. This reduces the uncertainty of the unchosen option because it's value is always 1 minus the value of the chosen option.

Thus, we next asked which model better explained matching law task behavior. We found that foraging was also a better fit to human behavior than RL in this new task (Figure 5B-C, FOR: log-likelihood = -6819, RL:log-likelihood=-7241 and relative AIC weight $< 5.1 * 10^{-32}$). (Note that we were unable to fit the model in 1/95 participants because that participant repeatedly selected the same option 300 times ([1.1%]; a similar proportion to the number that did so in Experiment 1: 4/258 [1.5%]).) To understand why foraging was a better fit, we generated an RL and foraging agent from each participant's parameters (see Methods) and then compared the simulated agents' behaviors with the humans' choice patterns. In line with Experiment 1, foraging better accounted for the participants' tendency to switch between options (Figure 5D-E, RL: MSSE = 0.38, foraging: MSSE = 0.06, Wilcoxon signed rank test $p < 1.5 * 10^{-7}$). Together, these results suggest that although a compare-alternatives algorithm could have offer an advantage in this task, the compare-to-threshold algorithm still better predicted participant behavior.

In Experiment 3, participants (270 people, 112 females, 2 other or non reporting) performed a 3-armed version of the bandit task in which we manipulated volatility (see Methods). Increasing the volatility of the environment increases uncertainty about the value of unchosen options because it increases the speed at which prior reward information becomes uninformative about the current state of the world. Moreover, prior research suggest that increasing volatility tends to increase switching behaviors in humans [37, 38], which could offer an advantage to RL, which is a less stable and persistent algorithm. Given that both Experiment 1 and 2 had similar low volatility (Figure 5F), we increased the volatility of the environment by low, medium, or high levels (Figure 5G).

We asked whether the foraging model fit better than the RL model despite manipulations of volatility. Regardless of the volatility of the environment, foraging was a better fit (Figure 5H-I, "high volatility" FOR: log-likelihood = -16,928 , RL:log-likelihood=-18,013 and relative AIC weight $< 1.2 * 10^{-32}$; "medium volatility" FOR: log-likelihood = -19,155, RL:log-likelihood=-19,955 and relative AIC weight $< 5.5 * 10^{-32}$; "low volatility" FOR: log-likelihood = -19,247 , RL:log-likelihood=-20,649 and relative AIC weight $< 1.8 * 10^{-32}$). Together, these results suggest that foraging better explained human behavior even in high volatility environments, regardless of how much information was available about the unchosen option.

Discussion

Even though the k-armed bandit is a classic testbed for RL algorithms [4, 6, 20, 30], and human and animal behavior on this task is often modeled with RL models [20, 21, 23, 30, 39], we found that human behavior was better described by a foraging algorithm that compared the value of an exploited option against a fixed threshold rather than comparing alternative options directly. Foraging was a better fit even when comparing alternatives would have been the more advantageous strategy but also was just consistently better across 3 variations of the task that manipulated the uncertainty of the unchosen option. Across these environments, the foraging models consistently better matched participants' tendency to switch options, both on average

and on an individual basis. The foraging models were even able to predict the existence of multiple participants who never switched at all: participants that essentially should not have existed under RL.

The observation that foraging models are a better fit to human behavior may superficially appear to contradict the large body of evidence for RL-like computations in the neuroscience literature. We know, for example, that the midbrain dopamine system calculates reward prediction errors that resemble those needed for reinforcement learning [40–47]. However, the foraging model introduced here is based on these same reward-prediction error calculations: this is precisely the mechanism that is used to track the value of the exploited option. This implies that neural evidence for a reward prediction error is not sufficient evidence for a traditional RL model. It could equivalently be interpreted as supporting the foraging model we develop here.

The key difference between the RL and foraging models is only that the former tracks and compares the values of a set of options, whereas the latter only tracks the value of one option. Further, the neural evidence that the brain tracks the values of all alternative options is not as strong as the evidence for reward prediction errors [20, 45, 48]. We would note that is especially important to be cautious in interpreting correlations between value and neural activity because (1) time series data are prone to spurious correlations, (2) putative neural correlates of value could be caused by other mechanisms, like mnemnonic errors [49], and (3) neural activity have a complex, nonlinear relationship with value even when value is explicitly cued [50]. RL models also do not naturally predict important nonlinearities in brain and behavior like abrupt state transitions in value-coding regions [21, 24, 51–53]. By contrast, the foraging model predicts that there should be nonlinearities in the neural and behavioral correlates of value whenever the threshold is crossed and exploration begins—exactly when such nonlinearities are observed experimentally [21, 24].

The RL models typically had fewer parameters than the foraging models, but they also kept track of more information than the foraging models. Even the best fitting foraging model only kept track of the current rate of reward of one option and how long it had been exploiting it. Conversely, the best fitting RL model kept track of the rewards and choice histories of every option it had sampled. Although there were only 2 options in Experiment 1 and the difference between tracking 2 and 4 variables seems trivial, this difference would quickly become substantial in tasks with more alternatives or continuous decision-spaces. Thus, while the RL models may be more parsimonious than foraging in that they have fewer parameters, they make less parsimonious assumptions about the information that must be tracked by the brain. In short, foraging was not only a better fit to behavior, but a more parsimonious explanation for it.

Ultimately, this paper reinforces and extends a body of recent observations that humans, like many other animals, can use foraging computations [15–17]. However, where previous studies of foraging computations in psychology and neuroscience have used foraging-specific tasks—tasks that are explicitly designed to encourage foraging strategies or computations [14–19, 54, 55]—the present study did the opposite. Through developing a new foraging model that is capable of navigating dynamic,

unpredictable environments, we determined that humans also prefer to use foraging computations even in tasks that were originally designed as RL testbeds.

Methods

Data Collection

Ethics Approval. All experiments procedures were in line with the standards set by the Declaration of Helsinki and were approved by the relevant ethical review boards (Experiment 1 and 2: the guidelines of the Comité d'Éthique de la Recherche en Sciences et Santé (CERSES) of the University of Montreal [Project #2021-1090], Experiment 3: Institutional Review Board of the University of Minnesota [STUDY00008486]).

Participants. Participants were recruited via the online platform, Amazon Mechanical (mTurk). To avoid bots and facilitate better data quality, participants were only accepted when they had a minimum of 5000 approved human intelligence tasks (HIT) and a minimal percentage of 98% in proportions of completed tasks that are approved by requesters. Geographical restrictions were set for US participants only. To help ensure that participants understood the task, there was an initial 25 trial block with fixed but randomly assigned reward probabilities (20%, 30% and 70%). To continue to the main task, participants were required to (1) switch between options at least twice, and (2) get reward more than 42% of the time during the practice block. The criteria and number of trials was chosen to ensure that people choosing at random would rarely pass through to to the main task. Participants were not allowed to repeat the experiment. All participants who successfully submitted the HIT, either the practice block or the full experiment, were paid a base rate of \$0.50, plus a bonus of \$3.61 mean $\pm$ \$0.97 SD based on their performance (for each trial that ended with a reward, participants were given a \$0.02 compensation). Participants provided written informed consent after the experimental procedure had been fully explained and were reminded of their right to withdraw at any time during the study.

For experiment 1 (Two-armed bandit task), a total of 258 participants (120 females, 137 males, 1 preferred not to say) completed the task. This dataset has been analyzed previously in a comparison of switching behavior between mice, monkeys, and humans [31], but all analyses presented here are new. For experiment 2 (Matching law task), a total of 95 participants (41 females, 51 males) completed the task. For experiment 3 (three-armed bandit task), a total of 270 participants (112 females, 156 males, 2 prefer not say) completed the task.

Experiment 1: Two-armed restless bandit task. To investigated how humans solved a restless bandit task, participants were required to repeatedly choose one of two options. Each option was associated with a probability of reward, which changed across trials and independently across options. The experiment consisted of 25 practice trials followed by 300 trials. To avoid environmental biases, each participant's reward schedule was randomly initialized. At any given step, the reward probability of each option could subsequently vary by a 0.1 increment with a hazard rate of 0.1, following a Bernoulli distribution.

Experiment 2: Matching law task. To understand how behavior changed under a different reward schedule, participants performed a modified version of the two-armed restless bandit task. Although both tasks are similar, they differ in how rewards are allocated to each of the two options. Specifically, in this task, the reward probabilities associated with each option varied inversely with respect to one another and across time.

Experiment 3: Volatile 3-armed bandit task. To explore how increased uncertainty in unchosen options influence behavior, participants performed a three-armed restless bandit task under various manipulations of volatility of the environment. This task differed from the two-armed restless bandit in two keys aspects. First, an additional option was introduced. This is common practice in the field, and previous studies show that increasing the number of options (up to 4) has no significant effect on the probabilities of switching and exploring [31]. Second, the change in reward probabilities was variable (probability of step = 1 or 0.1), with a different increment size for each participant (ranging from 0.05 to 0.667).

Since there were 2 manipulations in this task (i.e., probability of step and step size), volatility was assessed by looking at the auto-correlation of the reward schedule experienced by each participant. We estimated the rate of decay of the auto-correlation by measuring when it decayed to 0.75 of its initial value. We used this measurement as an arbitrary volatility index for each participant. We then divided the volatilities into 3 levels (high, medium, low).

Data Analysis

General Analysis Methods

Data was analyzed with custom software written in MATLAB and Python. Statistical tests were two-sided unless otherwise specified. Significance was compared against the standard alpha = 0.05 threshold throughout. Correlations are Pearson correlation unless otherwise indicated. A small number of participants (Experiment 1: 4/258 participants, 1.6% ; Experiment 2: 1/95, 0.95%) were excluded from some model-based analyses because they chose only one option in the main task, which made the model parameters unidentifiable. However, these participants were included in other analyses wherever it was possible to do so.

Foundations of Decision-Making Models

To determine whether participants were using foraging-like or RL-like computations, we fit a series of 10 cognitive models (5 RL models and 5 foraging models). All models were initialized with 50 random seeds and fit via maximum likelihood (minimizing the negative of the log likelihood; fminsearch, MATLAB, scipy.optimize.minimize, Python). Here, we will describe the standard formulations of both the RL and foraging models. The following section will explain the various extensions of each model.

All of the models, RL and foraging alike, had a classic delta-rule value updating process at their core,

$$v_{t+1} = v_t + \alpha(r_t - v_t) \ . \tag{1}$$

That is, values ($v_t + 1$) were updated at each time step t according to the difference between the observed reward (r_t) and the previous value (v_t; $v_0 = 0$ for all values and in all models). The magnitude of the update is scaled by a learning rate (α; constrained between 0 and 1 in all models). The delta-rule update dynamically calculates value as an exponentially-decaying recency-weighted average of reward. This is a canonical computation, with widespread neural and behavioral evidence supporting its existence in humans and other animals [1, 6, 20, 56–60].

Though both models have the delta-rule computation at their core, they differ in what the value represents. In the foraging model, the value is the value of a single focal option: it is the value of staying with or "exploiting" that option, so the update equation is precisely the one given above. In the RL model, there are multiple values because each option i has its own value and each is updated (or not) at each time step. One common approach, which we used in the standard RL model here, is to update only the value of chosen options,

$$v_{i,t+1} = \begin{cases} v_{i,t} + \alpha(r_t - v_{i,t}), & \text{if } i \text{ is chosen} \\ v_{i,t}, & \text{otherwise} \end{cases} . \tag{2}$$

A softmax function transformed values into choice probabilities in both the foraging and RL models. In RL models, the probability of choosing option i at time t was

$$p(\text{choice}_t = i) = \frac{e^{\beta v_{i,t}}}{\sum_{j=1}^{N} e^{\beta v_{j,t}}} , \tag{3}$$

where β is an inverse temperature parameter that controls the noise in the value comparison (constrained between 0 [high noise] and 100 [low noise] in all models). Because this task only had $N = 2$ options, we can expand the sum and rearrange this equation,

$$p(\text{choice}_t = i) = \frac{1}{1 + e^{-\beta(v_{i,t} - v_{j,t})}} , \tag{4}$$

which may clarify the point that this function compares the value of the two options in order to decide which to choose. In the foraging models, we used the same softmax equation, but replace the value of option i with the value of exploiting and the value of the alternative option with some threshold ρ. The threshold ρ could be thought of as the expected value of any alternative (the background rate of reward in the environment; [12] or as a literal threshold for exploration. This gives the probability of repeating, or exploiting, the current option as

$$p(\text{exploit}_t) = \frac{1}{1 + e^{-\beta(v_t - \rho)}} , \tag{5}$$

where $p(\text{explore}_t) = 1 - p(\text{exploit}_t)$. Because this model calculates the probability of a latent state (i.e., exploiting versus exploring), rather than a choice, we also need to specify choice probability distributions for each state. These were:

$$p(\text{choice}_t = i | \text{exploit}) = \begin{cases} 1, \text{if choice}_{t-1} = i \\ 0, \text{otherwise} \end{cases} , \text{and} \tag{6}$$

$$p(\text{choice}_t = i|\text{explore}) = \begin{cases} \frac{1}{N-1}, & \text{if choice}_{t-1} \neq i \\ 0, & \text{otherwise} \end{cases} . \tag{7}$$

Meaning that the model would repeat the previous choice when exploiting, but draw a choice randomly from the set of options excluding the previous choice when exploring. Because exploration ensured that a new option was chosen and value of exploiting it was unknown, the value of exploitation was reset to the threshold on each exploration trial.

Extensions to Decision-Making Models

In addition to the standard formulations of the RL and foraging models introduced above, we considered a variety of extensions that are commonly used to improve the fit of RL models. First, we introduced an asymmetrical learning parameter, meaning a way for models to learn at different rates from rewards and omissions. Adding asymmetry often improves the fit of RL models [23, 61] and it improved the fit of both models here. In both the RL asymmetry and foraging asymmetry models, the delta-rule update was rewritten

$$v_{t+1} = \begin{cases} v_t + \alpha_W(1 - v_t), & \text{if } r_t = 1 \\ v_t + \alpha_L(0 - v_t), & \text{if } r_t = 0 \end{cases} , \tag{8}$$

where v was the value of the chosen arm in the RL model or the value of exploitation in foraging. Both positive and negative learning rates were constrained between 0 and 1 in both models. Next, we introduced a decay parameter, meaning a way for some values to diminish over time without sampling. This approach is widely used in the reinforcement learning literature, where it has been interpreted as the indicative of limited memory resources. It consistently improves RL model fits [62–64] and improved the fit of both models here. In the reinforcement learning model, we implemented decay in the traditional way: instead of allowing the value of an unchosen option to carry forward unchanged in time, it decayed in proportion to delta (bounded at 0 and 1) as

$$v_{i,t+1} = \begin{cases} v_{i,t} + \alpha(r_t - v_{i,t}), & \text{if } i \text{ is chosen} \\ \delta v_{i,t}, & \text{otherwise} \end{cases} . \tag{9}$$

Although the foraging model has no unchosen values to decay, it does have a threshold parameter, ρ, which was previously modeled as static, but now allowed to change over time

$$p(\text{exploit}_t) = \frac{1}{1 + e^{-\beta(v_t - \rho_t)}} . \tag{10}$$

Specifically, ρ could now decay with each successive choice to the same option through updating it in a state-dependent way as a function of some decay parameter δ,

$$\rho_{t+1} = \begin{cases} \delta\rho_t, & \text{if exploit} \\ \rho_1, & \text{if explore} \end{cases} . \tag{11}$$

Finally, we considered 2 models that included history dependence, meaning an outcome-independent tendency to repeat past behavior. Adding dependence on previous choices tends to improve the fit of RL models [4, 23], likely because of the strong tendency towards hysteresis in the choices of humans and other animals. Adding history dependence also improved the fit of both models here. In reinforcement learning, one common approach adds a choice kernel (k) for each option i which is updated in parallel with the value as

$$k_{i,t+1} = \begin{cases} k_{i,t} + \alpha_c(1 - k_{i,t}), & \text{if choice} = i \\ k_{i,t} + \alpha_c(0 - k_{i,t}), & \text{otherwise} \end{cases} , \tag{12}$$

where α_c is the choice kernel learning rate parameter. The choice kernel value is then combined with value in the softmax decision rule,

$$p(\text{choice} = i) = \frac{e^{\beta(v_i + k_i)}}{\sum_{j=1}^{N} e^{\beta(v_j + k_j)}} . \tag{13}$$

Thus far, we have added only 1 choice-history dependent parameter to this RL model, the α_c parameter, and this model is the history kernel 1 RL model referred to in the text. However, it is common to give choice history its own inverse temperature parameter β_c . This is the history kernel 2 RL model,

$$p(\text{choice} = i) = \frac{e^{\beta v_i + \beta_c k_i}}{\sum_{j=1}^{N} e^{\beta v_j + \beta_k k_j}} . \tag{14}$$

Although there was no natural way to add a choice history kernel to the foraging model, we could use a nearly identical approach to add a state history kernel to the foraging model. This approach adds hysteresis at the level of the exploitation state, rather than at the level of the choices themselves. To do this, we introduced a state history kernel that was updated alongside the value of exploitation as

$$k_{t+1} = \begin{cases} k_t + \alpha_c(1 - k_{i,t}), & \text{if exploit}_t \\ 0, & \text{otherwise} \end{cases} . \tag{15}$$

Meaning that the state history kernel was re-set to 0 when the animal explored. In the history-kernel 1 foraging model, this kernel was then added directly to the exponential in the decision rule,

$$p(\text{exploit}_t) = \frac{1}{1 + e^{-\beta(v_t - \rho + k_t)}} . \tag{16}$$

In the history-kernel 2 foraging model, the kernel was inserted, but scaled with its own inverse noise parameter, β_k, as

$$p(\text{exploit}_t) = \frac{1}{1 + e^{-\beta(v_t - \rho) - \beta_k k_t}} . \tag{17}$$

Model Fitting Procedure

Maximum Likelihood Estimation (MLE) was performed to determine the parameter values of both standard models that most accurately described human behavioral data. The log likelihood (LL) was computed as follows:

$$LL = \log p(\text{choice}_{1:T}|\theta_m, m) = \sum_{t=1}^{T} \log p(\text{choice}_t, \theta_m, m) \ , \tag{18}$$

where $p(\text{choice}_{1:T}|\theta_m, m)$ is the probability of all choices (from 1 to T trials) given the parameters θ_m of each model m with $\theta_{RL} = [\alpha, \beta]$ and $\theta_{foraging} = [\alpha, \beta, \rho]$ for each participant. To find the maximum likelihood parameters, the negative log-Likelihood was minimized using an optimization function (Experiment 1: Matlab fminsearch; Experiment 2-3: Python scipy.optimize.minimize). The optimization process was initiated multiple times with randomly selected starting parameters to avoid local minima (Experiment 1: 20 times, Experiment 2-3: 50 times). The set of parameters with the lowest negative Log-Likelihood were selected. The Akaike Information Criteria was calculated to identify the best fitting model, adjusting for differences in the number of parameters, and Akaike weights were calculated to estimate the relative ability of each model to minimize information loss.

Simulations from Decision-Making Models

Performance Optimization. In order to determine if foraging had an advantage (or disadvantage) compared to RL in performance, we asked if the upper bound on performance was higher for foraging than it was for RL. This meant that we used simulation to identify the optimal parameter combination for each model in matched, randomly generated environments that matched the statistics of Experiment 1. We then compared the performance of these optimized models. Optimal parameters combination was selected so that it maximized the relative reward (minimizing the negative of the relative reward, Python scipy.optimize.minimize) while remaining within the participants' parameters range.

Cohort simulation. When simulating data from the standard foraging and RL models, our goal was to produce datasets under each model that would resemble the set of human participants as closely as possible. To do this, we took the stochastic reward schedule experienced by each participant and simulated a new sequence of choices in that reward environment from the RL and foraging models. The simulations were generated using fitted parameters from the participant who experienced that environment.

Exponential Mixture Model

We examined whether identifiable temporal patterns existed within the participants' choice sequences. If a single time constant (probability of switching) governed the behavior, we would expect to see exponentially distributed inter-switch intervals. That is, the distribution of inter-switch intervals should be well described by the following model:

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} \ , \tag{19}$$

where β is the survival parameter of the model (mean inter-switch interval). Although the time between switch decisions was largely monotonically decreasing and concave upwards, the distribution was not well described by a single exponential distribution. Therefore, we next fit mixtures of varying numbers of exponential distributions (1-4) in order to infer the number of switching regimes in these choice processes. For continuous-time processes, these mixture distributions would be of the form:

$$f(x) = \sum_{i=1}^{n} \pi_i e^{-\frac{x}{\beta_i}} \ , \tag{20}$$

where $1 \geq \pi_i \geq 0$ for all π_i, and $\sum_{i=1}^{n} \pi_i = 1$. Here, each β_i reflects the survival parameter (average inter-switch interval) for each component distribution i and the π_i reflects the relative weight of each component. Because trials were discrete, we fit the discrete analog of this distribution: mixtures of 1-4 discrete exponential (geometric) distributions [65]. Mixtures were fit via the expectation-maximisation algorithm and we used standard model comparison techniques [66] to determine the most probable number of mixing components. Two regimes (log-likelihood = -31,048, 3 parameter) were significantly better fit to data than one (log-likelihood = -34,645, 1 parameter, likelihood ratio test: ratio = 7193.0, p $< 10^{-32}$) and, while adding additional regimes continued to improve model fit (3 regimes: log-likehood = -30,893, 5 parameters; 4 regimes: log-likelihood = -30,884), improvement was already saturated at 2 regimes (Figure 5D).

Hidden Markov Model

To determine when models and participants were exploring (versus exploiting), we used a hidden Markov model (HMM; [21, 23, 31, 35, 36]. In an HMM, choices (y) are 'emissions' that are generated by an unobserved decision process that is in some latent, hidden state (z). Latent states are defined by both the probability of making each choice (k, out of Nk possible options), and by the probability of transitioning from each state to every other state. Our model consisted of two types of states, an explore state and the exploit states. The emissions model for the explore state was the maximum-entropy distribution for a categorical variable, a uniform distribution:

$$p(y_t = k | z_t = \text{explore}) = \frac{1}{N_k} \tag{21}$$

This means that we made the fewest number of assumptions possible about the choices that were made during exploration in order to avoid biasing the model towards or away from any particular type of policy. However, modeling exploratory choices with a uniform distribution does not imply, require, or enforce random decision-making during these states [35, 36]. Because exploitation involves repeated sampling of each option, exploit states only permitted choice emissions that matched one option. That is:

$$p(y_t = k|z_t = \text{exploit}_i, k = i) = 1$$
$$p(y_t = k|z_t = \text{exploit}_i, k \neq i) = 0 \tag{22}$$

The latent states in this model are Markovian, meaning that the current state (zt) depends only on the most recent state (zt-1):

$$p(z_t|z_{t-1}, y_{t-1}, ..., z_1, y_1) = p(z_t|z_{t-1}) \tag{23}$$

This means that we can describe the entire pattern of dynamics in terms of a time-invariant transition matrix between 3 possible states: three possible states (two exploit states and one explore state). This matrix is a system of stochastic equations that describe the one-time-step probability of transitioning between every combination of past and future states (i, j).

$$p(z_t = i|z_{t-1} = j) \tag{24}$$

Because we had only a limited number of trials for each participant (300), parameters were tied across exploit states: each exploit state had the same probability of beginning (from exploring) and of sustaining itself. For conceptual reasons, the model also assumed that participants started in exploration and had to pass through exploration in order to start exploiting a new option, even if only for a single trial [21, 23, 31, 35, 36]. We have previously shown that models that lack these constraints by design tend to approximate them when fit to sufficiently large datasets [21, 23].

Because the emissions model was fixed, certain parameters were tied, the structure of the transmission matrix was constrained, and the initial state was specified, the final HMM had only two free parameters: one corresponding to the probability of exploring, given exploration on the last trial, and one corresponding to the probability of exploiting, given exploitation on the last trial. We have previously reported that this constrained model does not underperform an unconstrained model [21, 23]. and that unconstrained models tend to closely resemble to the constrained model when fit to large amounts of data [21].

The HMM was fit via expectation-maximization using the Baum Welch algorithm [67]. This algorithm finds a (possibly local) maxima of the complete-data likelihood. The algorithm was reinitialized with random seeds 20 times, and the model that maximized the observed (incomplete) data log likelihood across all the sessions for each animal was ultimately taken as the best. To decode latent states from choices, we used the Viterbi algorithm to discover the most probable a posteriori sequence of latent states.

Explore/Exploit State Dynamics

In order to understand the dynamics of exploration and exploitation in the participants, RL models, and foraging models, we analysed the dynamics of the HMMs [23, 31, 36]. The fitted HMMs are a set of equations that describe probability of transitions between exploration and exploitation and vice versa, or the state dynamics of each agent. Methods from statistical thermodynamics can then be used to analyze these equations and generate insight into the potential energy of each state in each agent (Figure 4M).

In statistical thermodynamics, the potential energy of a state (E_i) is related to the long-time scale probability of observing a process (here, a decision-maker) in that state (p_i) via the Boltzman distribution,

$$p_i = \frac{1}{Z} e^{\frac{-E_i}{k_B T}} \ , \tag{25}$$

where Z is the partition function of the system, kB is the Boltzman constant, and T is the temperature. In a two-state system, the partition functions cancel out, the relative occupancy of the states is just a function of the difference in energy between them, and we can rearrange to express the difference in energy between two states as a function of the difference between them,

$$\ln\left(\frac{p_i}{p_j}\right) k_B T = E_j - E_i \ , \tag{26}$$

where $E_j - E_i$ is the difference in the energetic depth between the two states (i.e. the Gibbs Free Energy), which is proportional to the log odds of each state, up to some multiplicative factor, kBT.

To calculate the probability of exploration and exploitation (p_i and p_j), we solved for the stationary distribution π^* of the fitted HMMs, where π^* is the probability distribution that satisfies

$$\pi^* = \pi^* A_k \ , \tag{27}$$

where A_k is the transition matrix for agent k. If it exists, this distribution is a (normalised) left eigenvector of A_k with an eigenvalue of 1, so we solved for this eigenvector to determine the stationary distribution over explore and exploit states for each agent. We then took an average of these stationary distributions across all sessions for each species, and plugged these back into the Boltzman equations to calculate the relative energy (depth) of exploration and exploitation in (Figure 4M).

In order to calculate the height of the energetic barrier between exploration and exploitation, we built on the Arrhenius equation from chemical kinetics that relates the rate of transitions (k) between some pair of states to the activation energy required to affect these transitions (Ea):

$$k = C e^{\frac{E_a}{k_B T}} \ , \tag{28}$$

where C is a constant pre-exponential factor and kBT is again the product of temperature and the Boltzman constant. Rearranging to solve for activation energy yields our equation for the height of the energy barrier,

$$E_a = -\ln\left(\frac{k}{C}\right) k_B T \ , \tag{29}$$

which has an obvious similarity to the Boltzman distribution illustrated earlier, where the relative depth of each state was proportional to the probability of occupying each state and the activation energy is now proportional to the rate of transitions between states.

To create the dynamical landscape graphs (Figure 4M), transition matrices were calculated individually for all participants and for simulated data from both foraging

and RL algorithms. Energy measurements were then averaged within each class of agents. Note that our approach only identifies the energy of three discrete states (an explore state, an exploit state, and the peak of the barrier between them). These are illustrated by tracing a continuous potential through these three points to provide a physical intuition for the differences in explore/exploit dynamics between species.

Foraging Index

In order to determine if individual participants were more foraging-like (versus RL-like) in their approach to the task, we calculated the foraging index as the difference in the individual model likelihood (L) between the best-fitting foraging model (F*) and the best-fitting RL model (RL*) for each participant i:

$$\text{foraging index}_i = \mathcal{L}_{F^*,i} - \mathcal{L}_{RL^*,i} \tag{30}$$

Richness and Sparsity

The richness of the environment was defined as the sum over all the arms, while the sparsity of the environment was defined as the range over arms (i.e. the maximum difference), normalized by the sum. This normalization maximized the differences between the predictions of the compare-to-threshold and value-comparison strategies in Figure 3B-C. Although the sum of arms is approximately flat on average across all levels of the range (Figure 3A), there are more sums that can produce a narrow range than sums that can produce a wide range. Because the variance in the sum scales with range, a compare-to-threshold process could, in theory, have any relationship with difference, including a monotonically decreasing relationship that would closely resemble the value-comparison relationship for some threshold values. Normalizing the range by the sum when we calculate richness ensures that compare-to-threshold would produce a relationship between sparsity and exploration that is entirely distinct from the predictions of value-comparison. Following this normalization, value comparison still predicts a robust, monotonically decreasing relationship (Figure 3B), while compare-to-threshold predicts something between a monotonically increasing function and an inverted u-shape for most choices of parameters (Figure 3C).

Declaration of Interest

The authors have no competing interests to disclose.

Resource Availability

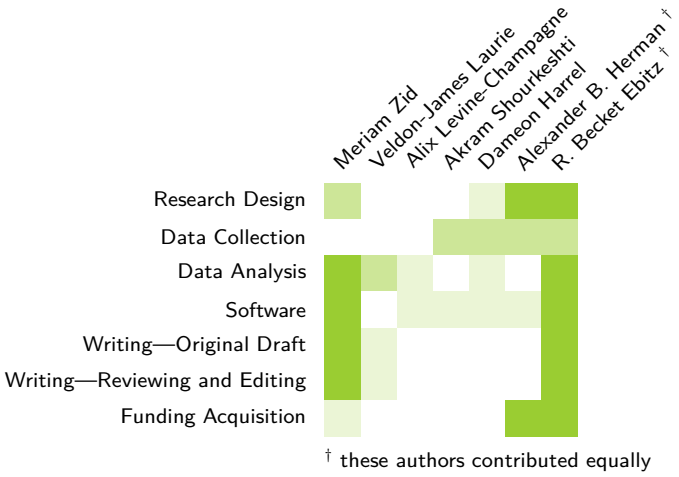
Material Availability

This study did not involve generating new material.

Data and Code Availability

Requests for Data and Code should be directed to and will be fulfilled by Lead Contacts. Data and code will be made publicly available upon acceptance of this article.

Author contributions



The author contribution matrix was adapted from [68].

Acknowledgements

The authors would like to thank Ossama Ghenissa, Rishabh Singhal, and Nicola Grissom for insightful discussion, Jorge Ramírez-Ruiz for feedback on an earlier draft of the manuscript, and Cathy Chen and Colin Meyer for invaluable technical support. This project was supported by an NSERC Discovery Grant (RGPIN-2020-05577; RBE), the Research Corporation for Science Advancement & Frederick Gardner Cottrell Foundation (Project #29087; RBE), the Canada Research Chair Dynamics of Cognition (FD507106; RBE), a Research Fellowship from the Jacobs Foundation (RBE), the Centre Interdisciplinaire de Recherche sur le Cerveau et l'Apprentissage (CIRCA; M.Z.), and the National Institutes of Health (R21 MH127607 and K23 MH050909; A.B.H.). The authors have no conflicts of interest to declare.

References

[1] Rescorla, R., Wagner, A.: A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Classical Conditioning II:

Current Research And Theory vol. Vol. 2, (1972). Journal Abbreviation: Classical Conditioning II: Current Research and Theory

[2] Levy, D.J., Glimcher, P.W.: The root of all value: a neural common currency for choice. Current Opinion in Neurobiology **22**(6), 1027–1038 (2012) https://doi.org/10.1016/j.conb.2012.06.001 . Accessed 2024-04-08

[3] McNamee, D., Rangel, A., O'Doherty, J.P.: Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. Nature Neuroscience **16**(4), 479–485 (2013) https://doi.org/10.1038/nn.3337 . Publisher: Nature Publishing Group. Accessed 2024-04-18

[4] Wilson, R.C., Collins, A.G.: Ten simple rules for the computational modeling of behavioral data. eLife **8**, 49547 (2019) https://doi.org/10.7554/eLife.49547 . Publisher: eLife Sciences Publications, Ltd. Accessed 2024-06-14

[5] Cisek, P., Hayden, B.Y.: Neuroscience needs evolution. Philosophical Transactions of the Royal Society B: Biological Sciences **377**(1844), 20200518 (2021) https://doi.org/10.1098/rstb.2020.0518 . Publisher: Royal Society. Accessed 2024-04-08

[6] Sutton, R.S., Barto, A.G.: Reinforcement Learning, Second Edition: An Introduction. MIT Press, ??? (2018). Google-Books-ID: uWV0DwAAQBAJ

[7] Wallis, J.D., Miller, E.K.: Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. European Journal of Neuroscience **18**(7), 2069–2081 (2003) https://doi.org/10.1046/j.1460-9568.2003.02922.x . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1460-9568.2003.02922.x. Accessed 2024-04-25

[8] Hosokawa, T., Kennerley, S.W., Sloan, J., Wallis, J.D.: Single-Neuron Mechanisms Underlying Cost-Benefit Analysis in Frontal Cortex. Journal of Neuroscience **33**(44), 17385–17397 (2013) https://doi.org/10.1523/JNEUROSCI.2221-13.2013 . Publisher: Society for Neuroscience Section: Articles. Accessed 2024-05-27

[9] Amemori, S., Amemori, K.-i., Yoshida, T., Papageorgiou, G.K., Xu, R., Shimazu, H., Desimone, R., Graybiel, A.M.: Microstimulation of primate neocortex targeting striosomes induces negative decision-making. European Journal of Neuroscience **51**(3), 731–741 (2020) https://doi.org/10.1111/ejn.14555 . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.14555. Accessed 2024-05-27

[10] Giarrocco, F., Costa, V.D., Basile, B.M., Pujara, M.S., Murray, E.A., Averbeck, B.B.: Motor System-Dependent Effects of Amygdala and Ventral Striatum Lesions on Explore–Exploit Behaviors. Journal of Neuroscience **44**(5) (2024) https://doi.org/10.1523/JNEUROSCI.1206-23.2023 . Publisher: Society for Neuroscience Section: Research Articles. Accessed 2024-05-27

[11] Charnov, E.L.: Optimal foraging, the marginal value theorem. Theoretical Population Biology **9**(2), 129–136 (1976) https://doi.org/10.1016/0040-5809(76)90040-X . Accessed 2024-04-30

[12] Stephens, D.W., Krebs, J.R.: Foraging Theory. Princeton University Press, ??? (1986)

[13] Gabay, A.S., Apps, M.A.J.: Foraging optimally in social neuroscience: computations and methodological considerations. Social Cognitive and Affective Neuroscience **16**(8), 782–794 (2021) https://doi.org/10.1093/scan/nsaa037 . Accessed 2024-06-06

[14] Hayden, B.Y., Pearson, J.M., Platt, M.L.: Neuronal basis of sequential foraging decisions in a patchy environment. Nature Neuroscience **14**(7), 933–939 (2011) https://doi.org/10.1038/nn.2856 . Publisher: Nature Publishing Group. Accessed 2024-06-12

[15] Kolling, N., Behrens, T.E.J., Mars, R.B., Rushworth, M.F.S.: Neural Mechanisms of Foraging. Science **336**(6077), 95–98 (2012) https://doi.org/10.1126/science.1216930 . Publisher: American Association for the Advancement of Science. Accessed 2024-06-06

[16] Wikenheiser, A.M., Stephens, D.W., Redish, A.D.: Subjective costs drive overly patient foraging strategies in rats on an intertemporal foraging task. Proceedings of the National Academy of Sciences **110**(20), 8308–8313 (2013) https://doi.org/10.1073/pnas.1220738110 . Publisher: Proceedings of the National Academy of Sciences. Accessed 2024-06-12

[17] Constantino, S.M., Daw, N.D.: Learning the opportunity cost of time in a patch-foraging task. Cognitive, Affective, & Behavioral Neuroscience **15**(4), 837–853 (2015) https://doi.org/10.3758/s13415-015-0350-y . Accessed 2023-12-19

[18] Kane, G.A., James, M.H., Shenhav, A., Daw, N.D., Cohen, J.D., Aston-Jones, G.: Rat Anterior Cingulate Cortex Continuously Signals Decision Variables in a Patch Foraging Task. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience **42**(29), 5730–5744 (2022) https://doi.org/10.1523/JNEUROSCI.1940-21.2022

[19] Blanchard, T.C., Hayden, B.Y.: Neurons in Dorsal Anterior Cingulate Cortex Signal Postdecisional Variables in a Foraging Task. Journal of Neuroscience **34**(2), 646–655 (2014) https://doi.org/10.1523/JNEUROSCI.3151-13.2014 . Publisher: Society for Neuroscience Section: Articles. Accessed 2024-06-14

[20] Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J.: Cortical substrates for exploratory decisions in humans. Nature **441**(7095), 876–879 (2006) https://doi.org/10.1038/nature04766 . Accessed 2024-06-12

[21] Ebitz, R.B., Albarran, E., Moore, T.: Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in Prefrontal Cortex. Neuron **97**(2), 450–4619 (2018) https://doi.org/10.1016/j.neuron.2017.12.007

[22] Pearson, J.M., Hayden, B.Y., Raghavachari, S., Platt, M.L.: Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multi-option choice task. Current biology : CB **19**(18), 1532–1537 (2009) https://doi.org/10.1016/j.cub.2009.07.048 . Accessed 2024-05-16

[23] Chen, C.S., Knep, E., Han, A., Ebitz, R.B., Grissom, N.M.: Sex differences in learning from exploration. eLife **10**, 69748 (2021) https://doi.org/10.7554/eLife.69748 . Publisher: eLife Sciences Publications, Ltd. Accessed 2024-05-16

[24] Shourkeshti, A., Marrocco, G., Jurewicz, K., Moore, T., Ebitz, R.B.: Pupil size predicts the onset of exploration in brain and behavior. bioRxiv, 2023–0524541981 (2023) https://doi.org/10.1101/2023.05.24.541981 . Accessed 2024-07-08

[25] 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. Philosophical Transactions of the Royal Society B: Biological Sciences **362**(1481), 933–942 (2007) https://doi.org/10.1098/rstb.2007.2098 . Accessed 2024-06-12

[26] Shteingart, H., Loewenstein, Y.: Reinforcement learning and human behavior. Current Opinion in Neurobiology **25**, 93–98 (2014) https://doi.org/10.1016/j.conb.2013.12.004 . Accessed 2024-06-12

[27] Palminteri, S., Khamassi, M., Joffily, M., Coricelli, G.: Contextual modulation of value signals in reward and punishment learning. Nature Communications **6**, 8096 (2015) https://doi.org/10.1038/ncomms9096

[28] Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., Palminteri, S.: Behavioural and neural characterization of optimistic reinforcement learning. Nature Human Behaviour **1**, 0067 (2017) https://doi.org/10.1038/s41562-017-0067

[29] Kennerley, S.W., Walton, M.E., Behrens, T.E.J., Buckley, M.J., Rushworth, M.F.S.: Optimal decision making and the anterior cingulate cortex. Nature Neuroscience **9**(7), 940–947 (2006) https://doi.org/10.1038/nn1724 . Publisher: Nature Publishing Group. Accessed 2024-06-13

[30] Guo, D., Yu, A.: Human learning and decision-making in the bandit task: Three wrongs make a right. In: 2019 Conference on Cognitive Computational Neuroscience. Cognitive Computational Neuroscience, Berlin, Germany (2019). https://doi.org/10.32470/CCN.2019.1356-0 . https://ccneuro.org/2019/Papers/ViewPapers.asp?PaperNum=1356 Accessed 2024-05-13

[31] Laurie, V.-J., Shourkeshti, A., Chen, C.S., Herman, A.B., Grissom, N.M.,

Ebitz, R.B.: Persistent Decision-Making in Mice, Monkeys, and Humans. bioRxiv. Pages: 2024.05.07.592970 Section: New Results (2024). https://doi.org/10.1101/2024.05.07.592970 . https://www.biorxiv.org/content/10.1101/2024.05.07.592970v1 Accessed 2024-05-15

[32] Katahira, K.: The statistical structures of reinforcement learning with asymmetric value updates. Journal of Mathematical Psychology **87**, 31–45 (2018) https://doi.org/10.1016/j.jmp.2018.09.002 . Accessed 2024-06-14

[33] Palminteri, S.: Choice-confirmation bias and gradual perseveration in human reinforcement learning. Behavioral Neuroscience **137**(1), 78–88 (2023) https://doi.org/10.1037/bne0000541 . Place: US Publisher: American Psychological Association

[34] Wilson, R.C., Bonawitz, E., Costa, V.D., Ebitz, R.B.: Balancing exploration and exploitation with information and randomization. Current Opinion in Behavioral Sciences **38**, 49–56 (2021) https://doi.org/10.1016/j.cobeha.2020.10.001 . Accessed 2024-06-05

[35] Ebitz, R.B., Tu, J.C., Hayden, B.Y.: Rules warp feature encoding in decision-making circuits. PLOS Biology **18**(11), 3000951 (2020) https://doi.org/10.1371/journal.pbio.3000951 . Publisher: Public Library of Science. Accessed 2024-05-17

[36] Ebitz, R.B., Sleezer, B.J., Jedema, H.P., Bradberry, C.W., Hayden, B.Y.: Tonic exploration governs both flexibility and lapses. PLOS Computational Biology **15**(11), 1007475 (2019) https://doi.org/10.1371/journal.pcbi.1007475 . Publisher: Public Library of Science. Accessed 2024-05-17

[37] Rushworth, M.F.S., Behrens, T.E.J.: Choice, uncertainty and value in prefrontal and cingulate cortex. Nature Neuroscience **11**(4), 389–397 (2008) https://doi.org/10.1038/nn2066 . Publisher: Nature Publishing Group. Accessed 2024-06-12

[38] Deserno, L., Boehme, R., Mathys, C., Katthagen, T., Kaminski, J., Stephan, K.E., Heinz, A., Schlagenhauf, F.: Volatility Estimates Increase Choice Switching and Relate to Prefrontal Activity in Schizophrenia. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging **5**(2), 173–183 (2020) https://doi.org/10.1016/j.bpsc.2019.10.007 . Accessed 2024-06-12

[39] Pearson, J.M., Heilbronner, S.R., Barack, D.L., Hayden, B.Y., Platt, M.L.: Posterior cingulate cortex: adapting behavior to a changing world. Trends in Cognitive Sciences **15**(4), 143–151 (2011) https://doi.org/10.1016/j.tics.2011.02.002 . Accessed 2024-06-13

[40] Montague, P.R., Dayan, P., Sejnowski, T.J.: A framework for mesencephalic dopamine systems based on predictive Hebbian learning. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience **16**(5), 1936–1947 (1996) https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996

[41] Schultz, W., Dayan, P., Montague, P.R.: A neural substrate of prediction and reward. Science (New York, N.Y.) **275**(5306), 1593–1599 (1997) https://doi.org/10.1126/science.275.5306.1593

[42] Montague, P.R., Berns, G.S.: Neural economics and the biological substrates of valuation. Neuron **36**(2), 265–284 (2002) https://doi.org/10.1016/s0896-6273(02)00974-1

[43] Cohen, M.X., Ranganath, C.: Reinforcement Learning Signals Predict Future Decisions. Journal of Neuroscience **27**(2), 371–378 (2007) https://doi.org/10.1523/JNEUROSCI.4421-06.2007 . Publisher: Society for Neuroscience Section: Articles. Accessed 2024-05-17

[44] Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., Uchida, N.: Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature **482**(7383), 85–88 (2012) https://doi.org/10.1038/nature10754 . Publisher: Nature Publishing Group. Accessed 2024-05-27

[45] Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., Berke, J.D.: Mesolimbic dopamine signals the value of work. Nature Neuroscience **19**(1), 117–126 (2016) https://doi.org/10.1038/nn.4173 . Publisher: Nature Publishing Group. Accessed 2024-05-27

[46] Cox, J., Witten, I.B.: Striatal circuits for reward learning and decision-making. Nature Reviews Neuroscience **20**(8), 482–494 (2019) https://doi.org/10.1038/s41583-019-0189-2 . Publisher: Nature Publishing Group. Accessed 2024-05-27

[47] Lee, R.S., Sagiv, Y., Engelhard, B., Witten, I.B., Daw, N.D.: A feature-specific prediction error model explains dopaminergic heterogeneity. bioRxiv. Pages: 2022.02.28.482379 Section: New Results (2024). https://doi.org/10.1101/2022.02.28.482379 . https://www.biorxiv.org/content/10.1101/2022.02.28.482379v3 Accessed 2024-06-14

[48] Tanaka, S.C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., Yamawaki, S.: Prediction of Immediate and Future Rewards Differentially Recruits Cortico-Basal Ganglia Loops. In: Ikeda, S., Kato, H.K., Ohtake, F., Tsutsui, Y. (eds.) Behavioral Economics of Preferences, Choices, and Happiness, pp. 593–616. Springer, Tokyo (2016). https://doi.org/10.1007/978-4-431-55402-8_22 . https://doi.org/10.1007/978-4-431-55402-8_22 Accessed 2024-05-27

[49] Ramírez-Ruiz, J., Ebitz, R.B.: "Value" emerges from imperfect memory. bioRxiv. Pages: 2024.05.26.595970 Section: New Results (2024). https://doi.org/10.1101/2024.05.26.595970 . https://www.biorxiv.org/content/10.1101/2024.05.26.595970v1 Accessed 2024-07-08

[50] Jurewicz, K., Sleezer, B.J., Mehta, P.S., Hayden, B.Y., Ebitz, R.B.: Irrational choices via a curvilinear representational geometry for value. bioRxiv. Pages: 2022.03.31.486635 Section: New Results (2022). https://doi.org/10.1101/2022.03.31.486635 . https://www.biorxiv.org/content/10.1101/2022.03.31.486635v1 Accessed 2024-04-29

[51] Durstewitz, D., Vittoz, N.M., Floresco, S.B., Seamans, J.K.: Abrupt Transitions between Prefrontal Neural Ensemble States Accompany Behavioral Transitions during Rule Learning. Neuron **66**(3), 438–448 (2010) https://doi.org/10.1016/j.neuron.2010.03.029 . Accessed 2024-04-29

[52] Karlsson, M.P., Tervo, D.G.R., Karpova, A.Y.: Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. Science (New York, N.Y.) **338**(6103), 135–139 (2012) https://doi.org/10.1126/science.1226518

[53] Pulcu, E.: A nonlinear relationship between prediction errors and learning rates in human reinforcement learning. bioRxiv. Pages: 751222 Section: New Results (2019). https://doi.org/10.1101/751222 . https://www.biorxiv.org/content/10.1101/751222v2 Accessed 2024-04-29

[54] Azab, H., Hayden, B.Y.: Correlates of economic decisions in the dorsal and subgenual anterior cingulate cortices. European Journal of Neuroscience **47**(8), 979–993 (2018) https://doi.org/10.1111/ejn.13865 . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.13865. Accessed 2024-04-18

[55] Azab, H., Hayden, B.Y.: Partial integration of the components of value in anterior cingulate cortex. Behavioral Neuroscience **134**(4), 296–308 (2020) https://doi.org/10.1037/bne0000382 . Accessed 2024-04-18

[56] Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning **8**(3), 229–256 (1992) https://doi.org/10.1007/BF00992696 . Accessed 2024-06-13

[57] Jacobs, R.A.: Increased rates of convergence through learning rate adaptation. Neural Networks **1**(4), 295–307 (1988) https://doi.org/10.1016/0893-6080(88)90003-2 . Accessed 2024-06-13

[58] Gluck, M.A., Bower, G.H.: From conditioning to category learning: an adaptive network model. Journal of Experimental Psychology. General **117**(3), 227–247 (1988) https://doi.org/10.1037//0096-3445.117.3.227

[59] Rumelhart, D.E., McClelland, J.L., Group, P.R.: Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations. The MIT Press, ??? (1986). https://doi.org/10.7551/mitpress/5236.001.0001 . https://direct.mit.edu/books/monograph/4424/Parallel-Distributed-Processing-Volume Accessed 2024-06-13

[60] Busemeyer, J.R., Stout, J.C.: A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. Psychological Assessment **14**(3), 253–262 (2002) https://doi.org/10.1037/1040-3590.14.3.253 . Place: US Publisher: American Psychological Association

[61] Ciranka, S., Linde-Domingo, J., Padezhki, I., Wicharz, C., Wu, C.M., Spitzer, B.: Asymmetric reinforcement learning facilitates human inference of transitive relations. Nature Human Behaviour **6**(4), 555–564 (2022) https://doi.org/10.1038/s41562-021-01263-w . Publisher: Nature Publishing Group. Accessed 2024-06-13

[62] Patel, N., Acerbi, L., Pouget, A.: Dynamic allocation of limited memory resources in reinforcement learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 16948–16960. Curran Associates, Inc., ??? (2020). https://proceedings.neurips.cc/paper/2020/hash/c4fac8fb3c9e17a2f4553a001f631975-Abstract.html Accessed 2024-06-14

[63] Santoro, A., Frankland, P.W., Richards, B.A.: Memory Transformation Enhances Reinforcement Learning in Dynamic Environments. Journal of Neuroscience **36**(48), 12228–12242 (2016) https://doi.org/10.1523/JNEUROSCI.0763-16.2016 . Publisher: Society for Neuroscience Section: Research Articles. Accessed 2024-06-14

[64] Yoo, A.H., Collins, A.G.E.: How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. Journal of Cognitive Neuroscience **34**(4), 551–568 (2022) https://doi.org/10.1162/jocn_a_01808 . Accessed 2024-06-14

[65] Barger, K.J.-A.: Mixtures of Exponential Distributions to Describe the Distribution of Poisson Means in Estimating the Number of Unobserved Classes (2006). Accessed 2024-06-13

[66] Burnham, K.P., Anderson, D.R. (eds.): Model Selection and Multimodel Inference. Springer, New York, NY (2002). https://doi.org/10.1007/b97636 . http://link.springer.com/10.1007/b97636 Accessed 2024-06-13

[67] Bilmes, J.A.: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models **4**(510) (1998)

[68] O'Shea, D.J., Duncker, L., Goo, W., Sun, X., Vyas, S., Trautmann, E.M., Diester, I., Ramakrishnan, C., Deisseroth, K., Sahani, M., Shenoy, K.V.: Direct neural perturbations reveal a dynamical mechanism for robust computation. bioRxiv. Pages: 2022.12.16.520768 Section: New Results (2022). https://doi.org/10.1101/2022.12.16.520768 . https://www.biorxiv.org/content/10.1101/2022.12.16.520768v1 Accessed 2024-07-08