Humans integrate heuristics and Bayesian inference to efficiently explore under uncertainty

Jing-Jing Li (jl3676@berkeley.edu)
University of California, Berkeley

Connor Chen (connorchen@berkeley.edu)
University of California, Berkeley

Anne G.E. Collins (annecollins@berkeley.edu)
University of California, Berkeley

Abstract

Exploring the environment efficiently and exploiting learned information effectively are crucial to intelligent agent behavior. Prior work has shown that humans can manage the exploration-exploitation trade-off not only action-by-action, but also at the strategy or rule level, using a heuristic on rule certainty (Collins & Koechlin, 2012). We evaluated this theory on a partially observable rule-switching task and collected human behavioral data (n=112) on two task variants with different levels of rule complexity to test whether taxing cognitive resources impacts exploration heuristics. Our results replicated previous findings, showing that the model is robust to dynamically switching task structure and increased executive demands due to rule complexity. Additionally, we identified a novel meta-heuristic of using high-level rule structure to inform decision-making and computationally characterized its integration with Bayesian inference to support efficient exploration. Through modeling analyses, we show that increased demand on executive function might interfere with this meta-cognitive process.

Keywords: Decision making, Learning, Bayesian modeling, Computational modeling, Mathematical modeling

Introduction

How do humans efficiently budget limited cognitive resources to learn and explore under uncertainty? Previous research has shown that the human brain can strategically switch between exploration and exploitation in multi-armed bandit environments (Cohen et al., 2007), and that such exploration can happen not only at the single action level, but also at more abstract rule and strategy levels (Collins & Koechlin, 2012; Donoso et al., 2014). Successful exploitation is highly desirable, as it can simultaneously maximize reward and cognitive efficiency. However, starting to exploit a strategy prematurely (without a sufficiently good policy) could lead to undesirable rewards. Thus, knowing when to switch from exploration to exploitation is crucial to cost-effective learning and decision-making.

Most research in the current literature has focused on strategic exploration and exploitation at the single action level. Many heuristic strategies for exploring the action space have been formalized (Schulz & Gershman, 2019), including random exploration (Wilson et al., 2014), ε-greedy (Sutton & Barto, 2018), softmax exploration (Cohen et al., 2007), and combinations thereof (Nassar & Frank, 2016). When the value or payoff of some action is uncertain or variable, uncertainty-driven exploration, which favors actions with less predictable values, such as Thompson sampling (Thompson, 1933) and the Upper Confidence Bound algorithm (Auer, 2002), can be beneficial. These algorithms may not be optimal or reward-maximizing, but they serve to help the agent gain information about action values and environment structures that could lead to more future rewards.

However, exploration at the individual action level alone suffers from the curse of dimensionality, as it does not scale efficiently to more complex environments. In tasks with complex and dynamic structures, which resemble real-life environments, humans learn hierarchically represented policies that abstract over sequences of actions (Botvinick et al., 2009; Xia & Collins, 2021; Li et al., 2022; Li & Collins, 2025). Algorithms that simply explore at the action level may fail in such environments as they are unable to attribute feedback to higher-level policies. Exploring at the strategy, task-set, or rule level may allow humans to efficiently generalize past knowledge to new contexts (Collins & Frank, 2013; Collins & Koechlin, 2012). Despite the rich literature on exploration strategies and the exploration-exploitation dilemma at the action level, our understanding of how humans solve these problems at the rule level remains limited. Previous work has shown that the exploration-exploitation trade-off between rules can be modeled by a heuristic policy, in which the agent switches from exploring to exploiting if they are sufficiently confident that they have learned the correct reward function in a temporarily stable environment (Collins & Koechlin, 2012; Donoso et al., 2014). Here, we extend this theory to a more complex environment with varying decision rule complexity to test its robustness and generalizability.

Prior work on bounded rationality (Simon, 1955, 1972) and resource rational analysis (Lieder & Griffiths, 2017, 2020) has shown that humans operate under constraints of limited time and cognitive resources, and accounting for such resource constraints in models may give rise to a different understanding of human behavior and cognition. Based on this literature, we hypothesized that increased rule complexity may demand more executive functions and cognitive resources, and, therefore, impact exploration policies. To do so, we designed a novel experimental paradigm featuring six decision rules, mapped to four actions, in which the rewarded rule switched episodically. By varying the number of steps required to follow a decision rule, we tested the effect of rule complexity on the exploration-exploitation trade-off.

Previous work has also shown that humans can learn and infer latent information about environment structures and use

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

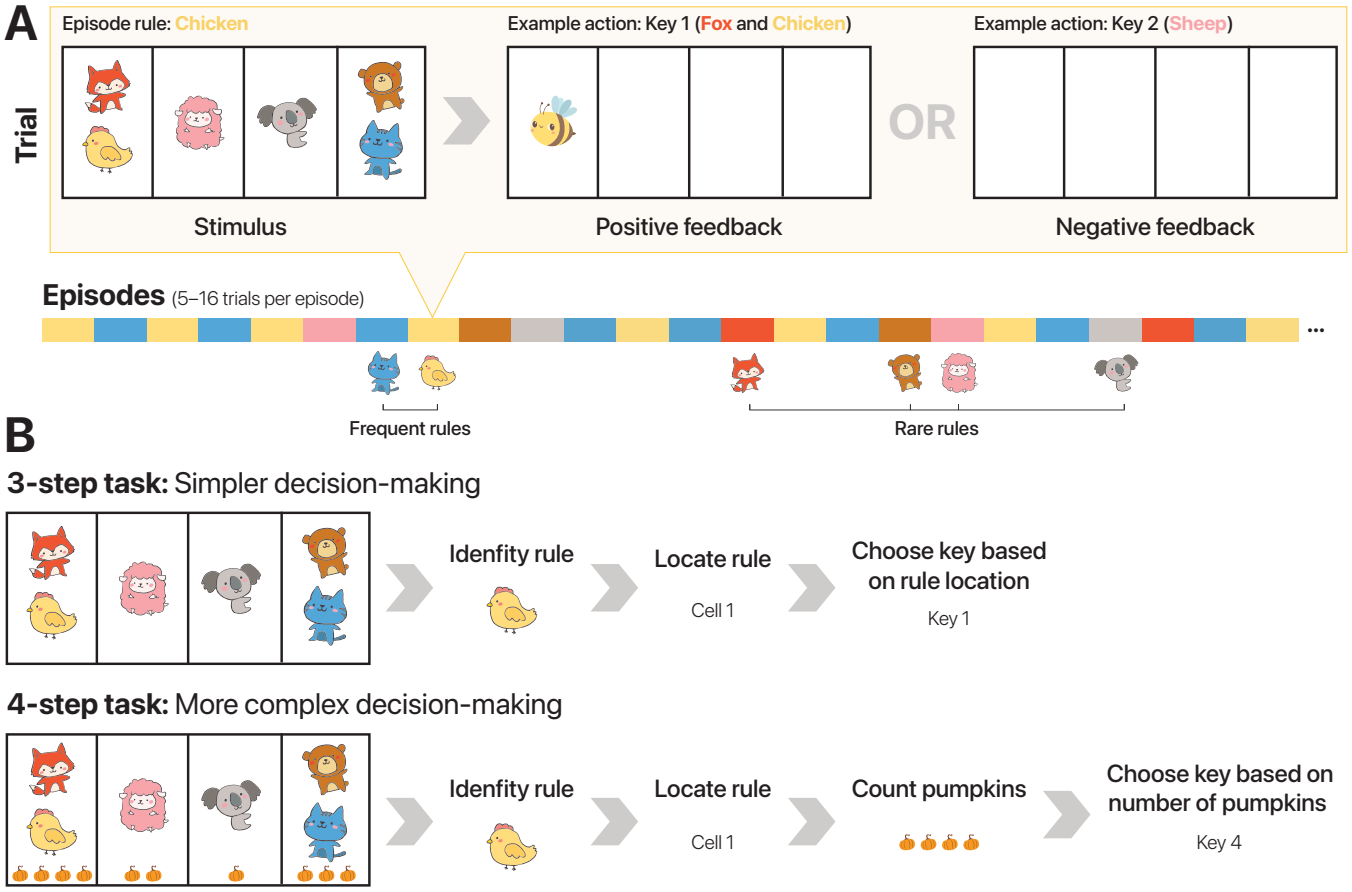

Figure 1: Experimental protocol. A: Participants played a hide-and-seek game of 720 trials to find a bee hiding behind one of six animals—the correct animal is referred to as the rewarded "rule," which switched every episode of 5-16 trials. Some rules were more frequently rewarded than others. On each trial, the six animals were grouped and placed in four cells. To choose an animal, the participant must select a key mapped to the cell containing it. B: Two versions of the task with different levels of complexity in decision-making were tested. In the 3-step task, the actions were mapped to the locations of the cells, while in the 4-step task, they corresponded to the number of pumpkins in each cell, adding an extra step of counting pumpkins to decision-making.

this knowledge to guide exploration (Daw & Shohamy, 2008; Collins & Frank, 2013). To investigate how humans integrate such meta-learned high-level environment structures into rule exploration, we further designed the task to reward some rules more frequently than others. Based on our human behavioral data, we formalized a novel cognitive model based on the PROBE model introduced by Collins & Koechlin (2012) to provide an algorithmic theory for this computational process.

Methods

Task

In the experiment, participants played a hide-and-seek game with the goal of finding a bee hiding behind one of six animals through trial and error (Figure 1). They were instructed that the bee might switch to a different animal unnoticed once it was found. Each participant completed 720 trials consisting of episodes of 5-16 trials in which the rewarded rule, represented by the animal the bee was hiding behind, switched between episodes, unannounced. To reduce the predictability of rule switches, they were designed to take place with a 25% probability once the participant had reached performance criteria (having completed at least 5 trials in the current episode with the last 3 trials correct). To test whether participants could acquire higher-order task structure through meta-learning, we pseudo-randomized the assignment of target rules to episodes such that two rules were four times more frequent than others.

On each trial, the six animals were arranged in groups of one or two and placed in four fields represented by horizontally aligned cells, which participants could choose from (Figure 1A). Mapping two animals to the same action introduced uncertainty in feedback credit assignment in addition to the uncertainty between the six possible rules—for example, in the trial illustrated in Figure 1A, if the participant's target rule

was the fox instead of the chicken and they made the choice corresponding to the left-most field (fox and chicken), they would be rewarded since the field also contains the correct rule (chicken), even if they internally selected an incorrect rule (fox). Between trials, the animals were shuffled and regrouped, such that the mapping between animals and actions changed every trial and must be reconsidered at each decision. Once an action was selected, the participant received positive feedback if the chosen cell contained the correct animal and negative feedback otherwise. On each trial, stimuli were displayed on the screen for up to 4 seconds until the participant selected an action. As soon as a response was made or 4 seconds had elapsed, feedback was displayed for 700 milliseconds, after which the next trial started following an 800-millisecond inter-trial interval.

To test the effect of decision-making complexity on participants' exploration policies, we designed two versions of the task (Figure 1B). In the simpler 3-step task, actions corresponded to the locations of the cells. To reach a decision, the participant needed to first identify a rule, then locate the cell the animal was in, and finally choose an action based on the location of the cell. In the more complex 4-step task, actions were mapped to the number of pumpkins in each cell (a unique number between 1 and 4) requiring the additional step of counting pumpkins once a cell was identified before choosing an action.

Participants

Human data collection for this research was approved by the Institutional Review Board of the University of California, Berkeley. Participants were recruited from Prolific within the United States and compensated at 12 U.S. dollars per hour. 56 and 56 participants completed the 3-step and 4-step tasks, respectively. After performance-based inclusion of participants with average pre-switch accuracy > 0.8, 50 and 44 participants remained in our analyses of both tasks.

Modeling

We tested four cognitive models that implement different combinations of Bayesian inference and heuristic policies, outlined below. Models were fit by maximum likelihood estimation over the log-likelihood objective, compared using the Akaike information criterion (AIC) (Akaike, 1974), and validated by simulating behavior with best-fit parameters. We followed modeling best practices as highlighted by Wilson & Collins (2019).

All models track the belief probability of each rule i for $i \in \{1,2,...,6\}$ (corresponding to the six animal rules), or the probability that rule i is the participant's target rule on trial t given the history of stimuli, actions, and rewards up to trial $t-1$, denoted for simplicity as $\Pr(\text{rule}_i\,;\,t)$. This belief probability is updated based on the action and reward information every trial. The rule-switching probability is integrated into the belief probabilities, such that the effective $\Pr(\text{rule}_i\,;\,t+1)$ is calculated by:

$$\Pr(\text{rule}_i\,;\,t+1) = \Pr(\text{rule}_i\,;\,t)\cdot(1-p_\text{s}) + \frac{1-\Pr(\text{rule}_i\,;\,t)}{Z}\cdot p_\text{s},$$

where Z is a normalizing factor to ensure the sum of all belief probabilities is 1 and p_s is the expected probability of a rule-switch, fixed to the ground truth value of $\frac{1}{7}$, calculated by inverting the minimum number of correct trials before a switch could occur (3 trials) plus the expected number of trials until a switch takes place with a 25% switch probability (4 trials). p_s approximates participants' average expectation of the frequency of rule switches. While p_s may be learned throughout the task, here we make the simplifying assumption of a fixed parameter value to minimize model complexity.

Bayesian model The Bayesian model fully tracks the probability of sampling each rule_i and updates it based on feedback trial by trial using Bayes' rule: the posterior $\Pr(\text{rule}_i\,;\,t+1)$ is proportional to

$$\Pr(\text{rule}_i\,;\,t)\cdot\Big((1-\varepsilon_B)\cdot\Pr(r_t \mid \text{rule}_i\,;\,t) + \varepsilon_B\cdot\frac{1}{n_A}\Big),$$

where $r_t \in \{0,1\}$ denotes the reward on trial t, $n_A = 4$ the number of available actions, and ε_B a free parameter for noise in likelihood computation, which helps model the strength of learning updates by modulating the impact of the likelihood vs. the prior. At decision time, the model chooses an action using a softmax policy over the log rule beliefs. $\Pr(a\,;\,t)$, the probability of sampling action a on trial t, is computed as:

$$\sum_{i:M(\text{rule}_i,t)=a}\frac{\exp\Big(\beta\cdot\log\big(\Pr(\text{rule}_i\,;\,t)\big)\Big)}{\sum_j\exp\Big(\beta\cdot\log\big(\Pr(\text{rule}_j\,;\,t)\big)\Big)}\cdot(1-\varepsilon)+\frac{1}{n_A}\cdot\varepsilon,$$

where $M(\text{rule}_i,t)$ returns the action mapped to rule_i on trial t, β is a free inverse temperature parameter, and ε adds uniform decision noise to the softmax policy. The ε-softmax policy enables parametric modeling of choice stochasticity due to uncertainty (β) and general decision noise (ε) (Nassar & Frank, 2016).

Meta-Bayesian model The meta-Bayesian model additionally meta-learns the frequency of each rule being rewarded over the course of the task and weights this information into the rule belief priors at the beginning of every trial, parameterized by w:

$$\Pr(\text{rule}_i\,;\,t) \leftarrow \Pr(\text{rule}_i\,;\,t)\cdot(1-w) + \frac{N^{t-1}_{\text{rule}_i}}{\sum_j N^{t-1}_{\text{rule}_j}}\cdot w,$$

where $N^{t-1}_{\text{rule}_i}$ denotes the number of times rule_i has been exploited up to trial $t-1$.

PROBE model The PROBE model, extended from Collins & Koechlin (2012), implements a heuristic that controls switches between latent exploration and exploitation states. The exploration phase policy and belief updating are identical to the Bayesian model, while the exploitation phase belief updates and policy differ: the policy always chooses the action according to the rule with the highest belief probability (the "default rule"). In the exploitation state, Bayesian updates are only applied to the exploited rule, while the other rules are assumed not to be tracked by the agent, such that their likelihoods are modeled randomly as $\frac{1}{n_A}$. This feature makes the PROBE model more resource-efficient and, there-

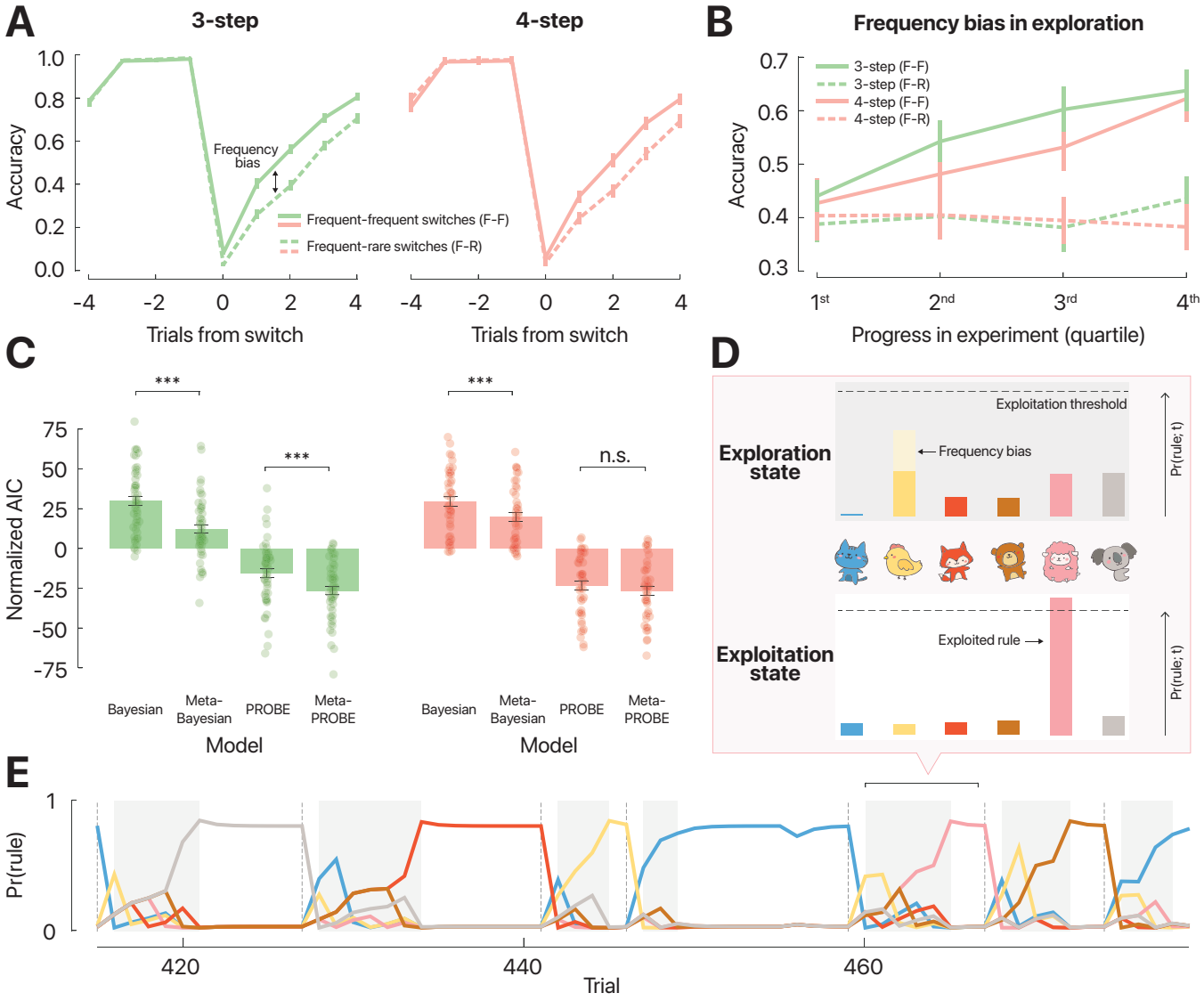

Figure 2: Human and model behavior. A: Human learning curves aligned to rule switches in the 3-step (n=50) and 4-step (n=44) task versions. Solid lines show the average human performance (measured by choice accuracy) around switches between frequent rules (F-F switches); dashed lines represent learning curves around switches from frequent rules to rare rules (F-R switches). B: Exploration accuracy (averaged over the first 3 trials post-switch) over the course of the experiment. Offsets in exploration accuracy between F-F and F-R switches reflect a bias for more frequent rules during exploration. C: The meta-PROBE model was the best fit model to human behavior as measured by AIC. D: Illustration of the meta-PROBE model mechanisms, showing example rule belief probabilities in both exploration and exploitation states. E: Model-estimated rule belief probabilities over time for an example participant. Vertical dashed lines mark target rule switch trials; shaded areas represent exploration phases as estimated by the meta-PROBE model, highlighting when the participant switched between exploiting a rule and exploring the rule space. Relevant model estimated parameter values for the example participant are: $w = 0.85$ and $\tau = 0.73$.

fore, more biologically plausible than the Bayesian model. Switches between exploration and exploitation are controlled by a free parameter τ, the exploitation threshold—once the maximum rule belief probability exceeds τ, the model exploits the default rule; if all belief probabilities are less than τ, the model explores all possible rules. Intuitively, once the PROBE model identifies a rule that it is sufficiently confident in, it exploits this rule until it is no longer rewarding.

Meta-PROBE model Extending the PROBE model, the meta-PROBE model additionally weights learned rule frequencies into the priors like the meta-Bayesian model, but only at the beginning of each exploration phase.

Results

Overall, participants performed similarly on both task variants. Figure 2A shows the average learning curves around rule switches from a frequent rule to another frequent rule (F-F; solid lines) and from a frequent rule to a rare one (F-R; dashed lines). Switches out of rare episodes are omitted here due to the scarcity of rare-to-rare switches. Participants quickly learned new rules within 5 trials post-switch, and learning fully saturated by the end of the episodes. No significant differences were found in exploration accuracy (averaged between the first 3 trials post-switch) between task variants (t-test $t = 1.60$, $p = 0.113$ for F-F and $t = 1.19$, $p = 0.238$ for F-R). We observed a frequency bias in human behavior, as indicated by faster learning of frequent rules than rare rules. Further analysis showed that this frequency bias developed throughout the experiment: Figure 2B illustrates the exploration accuracy (averaged over the first 3 trials post-switch) after F-F and F-R switches over the course of the experiment (divided into quartiles), showing an increasing frequency bias, as indicated by the expanding gap between F-F and F-R learning curves. The frequency bias, quantified by the difference in post-switch accuracy between F-F and F-R switches, was not significantly different between the 3-step and 4-step task variants (t-test $t = 0.425$, $p = 0.672$).

Our modeling analyses showed that the meta-PROBE model fit best to human choices on both versions of the task among all models we compared (Figure 2C). The large differences in AIC between the Bayesian models and the PROBE-based models suggest that human policies aligned better with the more resource-efficient explore-exploit heuristic than fully attentive Bayesian inference. The meta-PROBE model had a significant advantage over the PROBE model only in the 3-step task (paired t-test $t = -3.66$, $p = 6.18 \times 10^{-4}$), but not the 4-step task ($t = -1.88$, $p = 0.0668$).

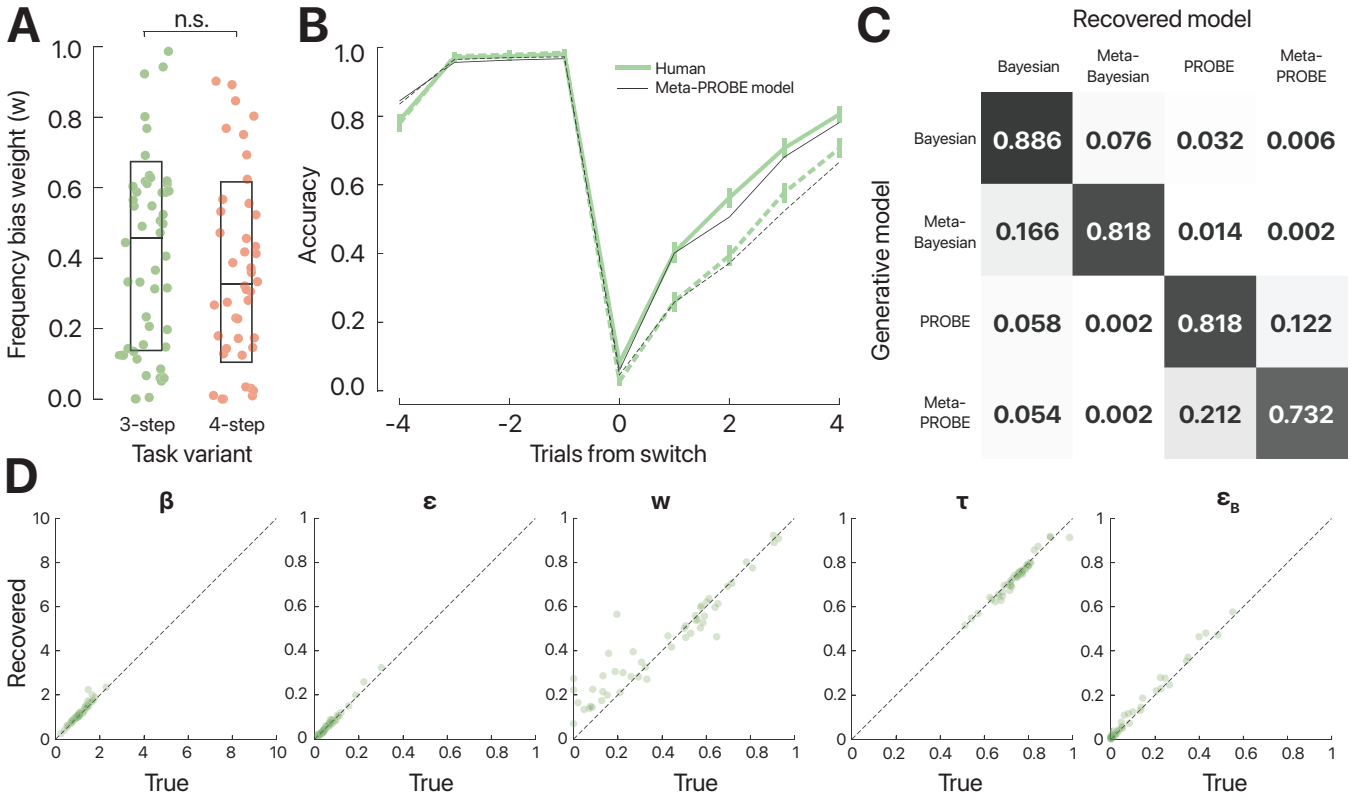

Figure 3: Interpretation and quality control of the meta-PROBE model fit. A: The individual fit frequency bias weight (w) parameter for both task variants. B: Validation of the meta-PROBE model against human behavior on the 3-step task. C: Model recovery results. D: Parameter recovery results.

However, the meta-Bayesian model fit significantly better than the Bayesian model for both task variants ($t = -5.36$, $p = 2 \times 10^{-6}$ for 3-step and $t = -5.12$, $p = 7 \times 10^{-6}$ for 4-step). This result indicates that the meta-learned rule frequency statistics played a significant role in informing exploration at both levels of rule complexity, despite a slightly more pronounced effect at lower rule complexity, which is consistent with the frequency bias observed in human behavior (Figure 2A, B).

In addition to group-level behavior, the frequency bias also manifested in the latent variables inferred by the meta-PROBE model based on individual behavior (Figure 2D, E). The meta-PROBE model incorporates the frequency bias into exploration by augmenting the prior of frequent rules (except for the last correct rule that has just stopped being rewarded) at the beginning of each exploration period (Figure 2D). This bias can continue to cause the agent to favor frequent rules over rare rules later in exploration—for example, around trials 427–429, 460–461, and 469–470 in Figure 2E, the meta-PROBE model inferred higher belief probabilities on frequent rules than rare rules, even though the rewarded rules were rare in those episodes.

The amounts of frequency bias in the 3-step and 4-step tasks were comparable—the estimated values of the frequency bias weight (w) parameter did not differ significantly between the two task variants (Figure 3A; t-test $t = 0.764$ and $p = 0.447$). The fit values of this parameter showed great individual variability. To ensure the validity of our model comparison and interpretation, we performed quality control analyses to show that our winning model (meta-PROBE) generated human-like behavior, was distinguishable from other models it was compared to, and had recoverable parameters. The meta-PROBE model validated well by closely reproducing qualitative human behavioral patterns, including the frequency bias (Figure 3B), which the PROBE model on its own could not reproduce in either task. Overall, the four models we compared could be robustly identified from one another in our model recovery analysis, in which data simulated by each model was used to find the best-fit model (Figure 3C). Meta-PROBE model parameters were highly recoverable—true generative parameter values and recovered parameter values through fitting were strongly and significantly correlated (Figure 3D; Spearman's $\rho = 0.98, p < 10^{-5}$ for β, $\rho = 0.99, p < 10^{-5}$ for ε, $\rho = 0.93, p < 10^{-5}$ for w, $\rho = 0.98, p < 10^{-5}$ for τ, and $\rho = 0.97, p < 10^{-5}$ for ε_B).

Discussion

In this work, we developed and tested a novel rule-switching behavioral paradigm featuring extended decision rules with statistical structures that could be meta-learned. Our results show that humans use a combination of heuristics and Bayesian inference to efficiently explore under uncertainty. In addition to replicating previous work in a more complex task (Collins & Koechlin, 2012), we characterized a novel meta-heuristic of using learned high-level reward structure (rule frequency statistics) to inform exploration by identifying a robust behavioral target (the frequency bias; Figure 2) that replicated between task variants with different levels of rule complexity. Our meta-PROBE model implements a plausible algorithm for this cognitive process, providing a computational account for how the meta-heuristic integrates with Bayesian inference to support efficient exploration.

Between the 3-step and 4-step tasks, we varied the complexity of decision rules to investigate the effects of cognitive load on exploration policies. We found that human behavior was qualitatively consistent despite increased rule complexity (Figure 2A, B). However, the slightly but not significantly smaller frequency bias (Figure 2B) and the marginal lack of significant advantage of the meta-PROBE model over the PROBE model ($p = 0.0668$) in the more cognitively demanding 4-step task suggest that increased cognitive load might interfere with the meta-heuristic and how it is integrated with Bayesian exploration (Figure 2C). This result suggests that increased demand on executive functions might slightly impair meta-learning under the constraint of limited cognitive resources. Nonetheless, the manipulation in the current experiment was insufficient to elicit such qualitative effects in human behavior, which necessitates further investigation with more pronounced differences in cognitive load between conditions or within-participant paradigms. While meta-PROBE implicitly incorporates cognitive load through the exploitation mechanism, alternative resource-rational models (Binz et al., 2024; Findling et al., 2019) may be tested to reveal additional computational accounts for cognitive processes engaged in this task.

Future work should investigate the neural mechanisms of heuristics and meta-learning in exploration under uncertainty. Prior research has shown that cortico-thalamic circuits play an important role in supporting cognitive flexibility in rodents and humans, and the thalamus can track frequent rules, which provides a compelling neural target (Rikhye et al., 2018; Chen et al., 2024). The behavioral paradigm developed in the current work can be extended to characterize the contributions of the thalamus to monitoring rule frequency and complexity. Additionally, further exploration of the model space using more advanced methods (Li et al., 2024; Pan et al., 2024) may help produce more informative latent state and variable inference for model-based analyses.

Acknowledgments

This research was supported by the National Institute of Mental Health under Award Number P50MH132642.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Auer, P. (2002). *Finite-time analysis of the multiarmed bandit problem.* Kluwer Academic Publishers.

Binz, M., Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., & Schulz, E. (2024). Meta-learned models of cognition. *Behavioral and Brain Sciences*, *47*, e147.

Botvinick, M. M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *cognition*, *113*(3), 262–280.

Chen, X., Leach, S. C., Hollis, J., Cellier, D., & Hwang, K. (2024). The thalamus encodes and updates context representations during hierarchical cognitive control. *PLoS biology*, *22*(12), e3002937.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). 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.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, *120*(1), 190.

Collins, A. G., & Koechlin, E. (2012). Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS biology*, *10*(3), e1001293.

Daw, N. D., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, *26*(5), 593–620.

Donoso, M., Collins, A. G., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486.

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature neuroscience*, *22*(12), 2066–2077.

Li, J.-J., & Collins, A. G. (2025). An algorithmic account for how humans efficiently learn, transfer, and compose hierarchically structured decision policies. *Cognition*, *254*, 105967.

Li, J.-J., Shi, C., Li, L., & Collins, A. G. (2024). Dynamic noise estimation: A generalized method for modeling noise fluctuations in decision-making. *Journal of Mathematical Psychology*, *119*, 102842.

Li, J.-J., Xia, L., Dong, F., & Collins, A. G. (2022). Credit assignment in hierarchical option transfer. In *Cogsci... annual conference of the cognitive science society. cognitive science society (us). conference* (Vol. 44, p. 948).

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological review*, *124*(6), 762.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Nassar, M. R., & Frank, M. J. (2016). Taming the beast: extracting generalizable knowledge from computational models of cognition. *Current opinion in behavioral sciences*, *11*, 49–54.

Pan, T.-F., Li, J.-J., Thompson, B., & Collins, A. G. (2024). Latent variable sequence identification for cognitive models with neural bayes estimation. *arXiv preprint arXiv:2406.14742*.

Rikhye, R. V., Gilra, A., & Halassa, M. M. (2018). Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nature neuroscience*, *21*(12), 1753–1763.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, *55*, 7–14.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 99–118.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization: A Volume in Honor of Jacob Marschak/North Holland*.

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

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3-4), 285–294.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*, *8*, e49547.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of experimental psychology: General*, *143*(6), 2074.

Xia, L., & Collins, A. G. (2021). Temporal and state abstractions for efficient learning, transfer, and composition in humans. *Psychological review*, *128*(4), 643.