Neural network modeling reveals diverse human exploration behaviors via state space analysis

Hua-Dong Xiong (hdx@arizona.edu)*
Department of Psychology, University of Arizona
Tucson, AZ, 85721, USA

Li Ji-An (jil095@ucsd.edu)*
Neurosciences Graduate Program, University of California San Diego
La Jolla, CA, 92093, USA

Marcelo G. Mattar (marcelo.mattar@nyu.edu)†
Department of Psychology, New York University
New York, NY, 10003, USA

Robert C. Wilson (bob@arizona.edu)†
Department of Psychology and Cognitive Science Program, University of Arizona
Tucson, AZ, 85721, USA

*Equal contribution.
†Corresponding author.

Abstract

The exploration-exploitation trade-off, balancing the acquisition of new information with the utilization of known resources, is a fundamental dilemma faced by all adaptive intelligence. Despite our understanding of models based on normative principles, the diverse explore-exploit behaviors of natural intelligence remain elusive. Here, using neural network behavioral modeling and state space analysis, we examined the diverse human exploration behaviors under a novel two-armed bandit task called Changing Bandit, designed to simulate real-world environmental volatility where exploration becomes essential. Examining behavior in the belief state space of this task, we characterized the disparities across artificial agents with decision boundaries. To extend this analysis to human data, a circumstance where choices are too sparse in the belief state space, we trained a recurrent neural network (RNN) model to predict humans' choices given past observations. This RNN model outperforms all existing cognitive models. Probing the RNN's decision boundaries, we found substantial individual differences that evade classical cognitive models. Additionally, our RNN revealed a tendency of "high-stay, low-shift" used by humans in response to higher environmental volatilities. Our work offers a promising approach for investigating diverse decision-making strategies in humans and animals.

Keywords: explore-exploit dilemma; recurrent neural network; state space analysis; computational modeling

Introduction

Exploration and exploitation are two widely studied elements in decision-making processes (Wilson, Geana, White, Ludvig, & Cohen, 2014; Wilson, Bonawitz, Costa, & Ebitz, 2021). However, the diversity in human exploration behavior remains challenging to characterize and comprehend. Here we present a novel approach combining neural network modeling and state space analysis to investigate human exploration in the Changing Bandit task. By examining decision boundaries of several artificial agents and a RNN model trained to predict human choices, we uncover numerous commonalities and individualities in human exploration.

Results

The Changing Bandit Task. In this task, agents make a series of choices between two options that payout different rewards (Fig. 1a). The reward from each option stays constant for several trials, then changes randomly, abruptly, and independently (with the hazard rate h in each trial) to a new value sampled from a uniform distribution (from 1 to 99 points). Unlike classic reversal learning tasks, a change in the reward of one option does not imply that the reward of the other option has changed. The longer agents exploit the same option, the more likely the other option has changed and the more uncertain they should be about its reward. Therefore, to maximize total rewards, the agent should constantly assess which option is better: exploiting the current known option or exploring the other more uncertain option.

State space analysis of the task. Optimal performance in the task (Fig. 1b) is achieved by solving Bellman's equation in a three-dimensional state space that captures the belief state of the task. The dimensions of this space are the reward from the option just exploited ("Current Reward"), the last reward from the other option ("Other Reward"), and the number of successive stay trials since last explored ("# Successive Stays", or "SS").

Optimal model-based agent. The optimal behavior is a deterministic function of the belief states. This agent explores or exploits depending on where the current state is relative to a Decision Boundary (DB) that cuts through the three-dimensional belief space (Fig. 1c, f). Benefiting from the encoding of the transition structure between belief states, it can gain information bonuses via exploration to maximize the future expected reward. Therefore, the DB of the model-based agent shows that the model will explore in some situations even in which the immediate expected payoff from exploring is less than exploiting (gray region in Fig. 1c).

Myopic model-free agent. This agent updates the chosen action value with the immediate reward and gradually forgets the unchosen action value. It is mathematically equivalent to a short-sighted model-based agent that only maximizes its immediate reward, not its long-term expected future reward. Thus, this model has a more conservative DB that does not favor exploration (Fig. 1c, f). The discrepancy in the DB between model-based and model-free models increases when the other reward is more uncertain (i.e., a larger SS).

Meta reinforcement learning (meta-RL) model. Since the meta-RL framework has been well-demonstrated to capture animals' reinforcement learning at both the behavioral and neural levels (Wang et al., 2018), we also analyze a meta-RL agent trained to maximize their task rewards. Our meta-RL agent has a GRU (gated recurrent unit) network with 100 hidden units, trained with the Advantage Actor-Critic (Wu, Mansimov, Grosse, Liao, & Ba, 2017). Our analysis shows that the meta-RL agent has near-optimal performance (Fig. 1b) but with a strategy substantially different from the optimal agent (Fig. 1f). As SS increases, it gains information bonuses by expanding the shift region.

RNN model reveals commonalities and individualities in human behavior. A total of 29 participants underwent 3 blocks of 1500 trials each, with one block dedicated to each hazard rate ($h = 0.1,\ 0.2,\ 0.4$), following a 300-trial practice session. Obtaining DBs for human behavior is challenging because the human choice data are too sparse in the belief state space. Thus, we train a GRU network (50 hidden units) to predict humans' behavior given history (Dezfouli, Griffiths, Ramos, Dayan, & Balleine, 2019), equipped with a subject embedding layer to capture individual differences (Song, Niv,

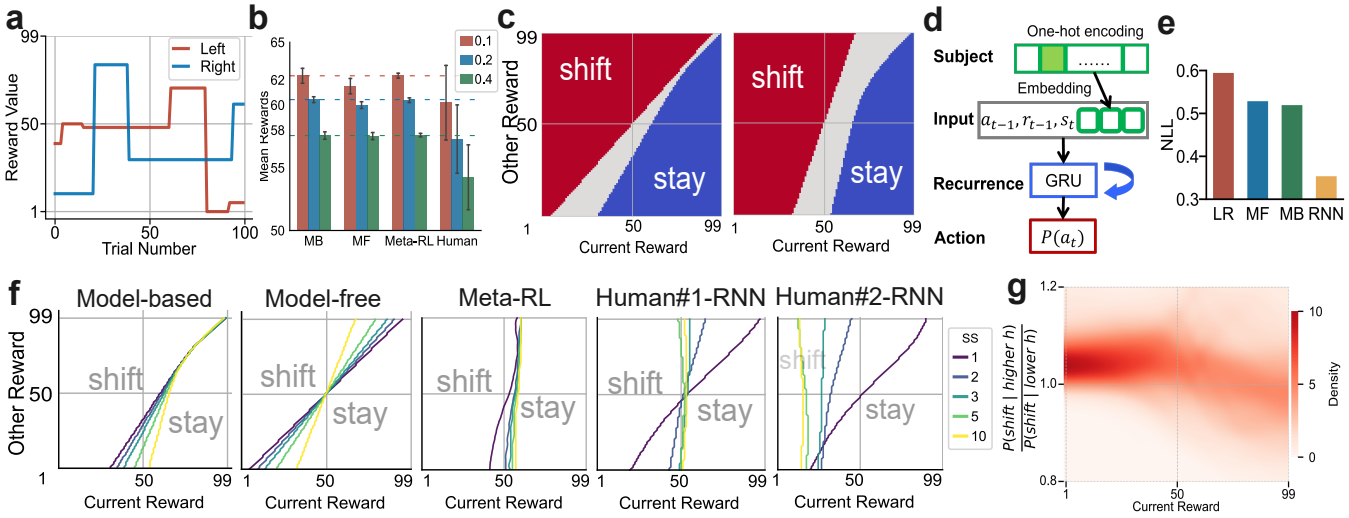

Figure 1: **a)**. Subjects choose between two options. Rewards from each option are constant for several trials but change randomly and independently in some trials (hazard rate h). **b)** Task performance of artificial agents and humans (MB: model-based; MF: model-free; Meta-RL: meta reinforcement learning) across different hazard rates. Dashed lines indicate the optimal performance. **c)** Comparison of the optimal model-based agent's decision boundaries (DBs; separating gray and blue) and the myopic model-free agent's DBs (separating gray and red) for $h = 0.1$. The two agents both shift (explore) in the red region and stay (exploit) in the blue region for SS = 1 (left) and SS = 10 (right). In the gray region, the model-based agent explores since it considers the expected future value, while the model-free agent exploits due to short-sightedness. **d)** The structure of the RNN model with subject embedding, receiving the current observation (s_t, the hazard rate), last-trial action a_{t-1}, last-trial reward r_{t-1}, and the current subject embedding as inputs. **e)** The negative log likelihood (NLL) of models fitted to human subjects (the lower the better). The RNN ourperforms existing cognitive models in predicting human choices given history. LR, logistic regression; MF, model-free; MB, model-based. **f)** The decision boundaries for different SS across agents ($h = 0.1$). **g)** The environmental volatility (hazard rate) influences shift/explore probability provided by the subject-simulating RNN.

& Cai, 2021). Unlike classical cognitive models, RNNs have a stronger capacity to characterize human decision-making processes without manually engineering the architecture or making explicit assumptions for underlying cognitive processes. Our RNN (Fig. 1d) outperms the best known cognitive models in predicting human choices (Fig. 1e, all models evaluated by negative log likelihood (NLL) via cross-validation).

Our RNN model's predictive performance substantially benefits from subject embedding (reducing NLL by 0.06 compared to a RNN without subject embedding). We found that the subject-prompted RNN's task performance strongly correlates with subjects' task performance ($r = 0.69, p < 10^{-4}$), and that principal components (PCs) of subject embeddings encode the subject's sensitivity (loadings in the logistic regression) to last-trial actions ($r = -0.62, p < 10^{-4}$ for PC1) and last-trial rewards ($r = 0.58, p < 10^{-4}$ for PC3). These results suggest that our RNN model with subject embedding has captured subject-specific patterns.

We then characterized subject-specific strategies using DBs provided by the RNN. For instance, in two example subjects' strategies (Fig. 1f, rightmost), we found that RNN-Subject-2 shows a DB similar to RNN-subject-1 for SS = 1, but a DB with narrower shift/exploration region than RNN-Subject-1 as SS increases. This result indicates that Subject 2 has a greater tendency to stay on the current action for larger SS.

Finally, we compared the subjects' tendency to explore (action probabilities provided by the RNN) in environments with low and high volatilities (Fig. 1g). We found that, when the current reward is low (i.e., below 50), subjects showed a tendency to explore (shift) more frequently in more volatile environments than in less volatile ones. As the current reward increases, this tendency gradually shifts towards exploitation (stay) on average, though with a growing variance). This finding suggests that our RNN can capture the tendency for humans to exhibit a "high-stay, low-shift" behavior in response to increasing environmental volatility, which classical models have failed to capture.

Conclusion

Characterizing behaviors in naturalistic sequential decision-making tasks is challenging, partially due to the complicated and often idiosyncratic dependency on prior experience. Traditional computational models usually fall short in capturing intricacies in humans' and animals' exploration behavior, neglecting important behavioral patterns. In contrast, our method effectively uncovers these nuanced exploration patterns. Our work contributes to the growing field of research on human and animal decision-making strategies and provides a promising approach for future studies of exploration behavior in natural and machine intelligence.

Acknowledgments

This work was funded by Kavli Institute for Brain and Mind Innovative Research Grant #2022-2209 to LJA; a National Institutes of Health grant R01AG061888 to RCW; a Psychology Department pilot grant at University of Arizona to RCW.

We thank the support from Swarma Club and AI + Science Reading Group supported by the Save 2050 Programme jointly sponsored by Swarma Club and X-Order.

References

Dezfouli, A., Griffiths, K., Ramos, F., Dayan, P., & Balleine, B. W. (2019). Models that learn how humans learn: The case of decision-making and its disorders. *PLoS computational biology*, *15*(6), e1006903.

Song, M., Niv, Y., & Cai, M. (2021). Using recurrent neural networks to understand human reward learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., ... Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, *21*(6), 860–868.

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current opinion in behavioral sciences*, *38*, 49–56.

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.

Wu, Y., Mansimov, E., Grosse, R. B., Liao, S., & Ba, J. (2017). Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. *Advances in neural information processing systems*, *30*.