Task-space dimensions guide human exploration in complex environments

Jiahui An
Chinese Institute for Brain Research, Beijing, China
Peking Union Medical College, Beijing, China

Jiewen Hu
Chinese Institute for Brain Research, Beijing, China
Peking Union Medical College, Beijing, China

Yilin Elaine Wu
China Institute of Brain Research, Beijing, China
Peking Union Medical College, Beijing, China
Nanjing University, Nanjing, China

Siyu Ning
Chinese Institute for Brain Research, Beijing, China
Peking Union Medical College, Beijing, China
China Agricultural University, Beijing, China

Fanyu Zhu
Chinese Institute for Brain Research, Beijing, China
Peking Union Medical College, Beijing, China

Yuhang Pan
Chinese Institute for Brain Research, Beijing, China
Peking Union Medical College, Beijing, China
Zhejiang University, Hangzhou, China

Ni Ji*
Chinese Institute for Brain Research, Beijing, China
Peking Union Medical College, Beijing, China

*corresponding author: **niji@cibr.ac.cn**

Abstract

Humans frequently make decisions in complex, high-dimensional environments, where identifying task-relevant information is critical for rapid behavior optimization. Humans outperform standard reinforcement learning agents in navigating such complexity, yet the cognitive strategies of humans remain unclear. To address this, we developed a novel multi-dimensional learning task in which only a subset of dimensions is reward-related. Crucially, unlike prior studies, subjects are uninformed of the true task dimensionality and have to identify them through exploration. This design closely mimics the ambiguity in real-world tasks. Our results have identified two stereotyped choice patterns that reveal "dimension-guided" strategies in exploration and exploitation. Cross-subject analyses suggest that dimension-guided exploration may promote the efficiency of reward-based learning. These findings indicate that humans leverage task dimensionality to guide exploration, and provide inspiration for improving exploration efficiency in AI agents.

Keywords: decision-making; multi-dimensional task; dimension-guided exploration; reward-based learning

Introduction

Humans constantly make decisions in complex environments where the outcome (or value) of one's action could potentially depend on a large array of environmental factors (i.e. dimensions). While the true action values may only depend on a few factors (i.e. the true dimensionality of the task is low), that information is not a priori known and has to be learned through exploratory interaction with the environment. Despite the demonstrated efficiency, how humans effectively explore and rapidly learn in high-dimensional task environments remains an open question.

In contextual bandit problems, previous studies have found that humans learn the mapping between contextual cues and action values in a way akin to Gaussian process regression (Krause & Ong, 2011; Schulz et al., 2018; Wu et al., 2018). In these studies, however, the dimensionality of the contextual space was often low ($\leq$2) and the number of reward-relevant dimensions explicitly conveyed to the subjects, reducing the need for active exploration and inference on task dimensionality (Lucas et al., 2015; Parpart et al., 2017).

Other studies examined value learning in multidimensional environments (Ballard et al., 2018; Mack et al., 2016; Niv et al., 2015; Wilson & Niv, 2012; Goodman et al., 2008; Wunderlich et al., 2011). They found that mechanisms such as Bayesian rule learning (Ballard et al., 2018), selective attention (Marković et al., 2015; Wilson & Niv, 2012) and serial hypothesis testing (Niv et al., 2015; Leong et al., 2017) could explain human choice behavior. However, since the subjects were pre-informed of the number of reward relevant dimensions (which was 1), it remained unclear how humans may explore in environments with unknown dimensionality and how they learn to identify the reward-relevant dimensions (Akaishi et al., 2016; Farashahi et al., 2017; Wang & Rehder, 2017).

Here we designed a novel behavior paradigm, where subjects explored a latent multi-dimensional environment and were told to arrange multiple options based on their estimated value. We found that subjects exhibited exploratory and exploitative choice patterns that organized around the true task dimensions. We further found that such dimension-guided exploration strategy correlate well with high task performance and rapid learning.

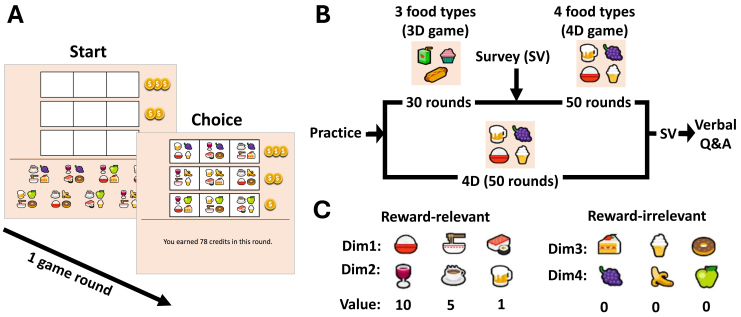

Figure 1: Task paradigm. **A)** Process of one game round. **B)** The entire experiment process. **C)** The score-relevant/irrelevant dimension settings for each participant (using 4D game as an example).

Method

Participants: 77 cognitively healthy adults (mean age = 26.6, ranging between 19-51 years; 70.1% females) were randomly assigned to two groups: the One-Game Group (n = 39) and the Two-Game Group (n = 38).

Procedures: As shown in **Figure 1B**, participants completed a practice session to learn basic game mechanics. The Two-Game Group subsequently completed a 30-round 3D (with this game dimension corresponding to food category) game followed by a 50-round 4D game. The One-Game Group directly engaged in the 4D game. After each game, participants completed a questionnaire assessing game engagement and learning outcomes. Finally, an oral interview gathered qualitative insights into their strategies and decision-making processes.

Task: In each game, participants were instructed to infer a customer's food preferences. They arranged 9 non-identical stimuli into three rows with instructed decreasing weights (top$>$middle$>$bottom, **Figure 1A**) in each round. A score feedback was given after each round's choice. The goal of the game was to obtain the maximum possible score in finite rounds. In each game session, two independent reward-relevant dimensions were randomly selected, within which each food item was assigned fixed values (**Figure 1C**). Participants were not informed of the number, identities of reward-relevant dimensions or the reward policy beforehand. The game ended automatically upon achieving the maximum possible score and proceeded to the Q&A phase.

Result

Humans exhibit two stereotyped and persistent choice patterns

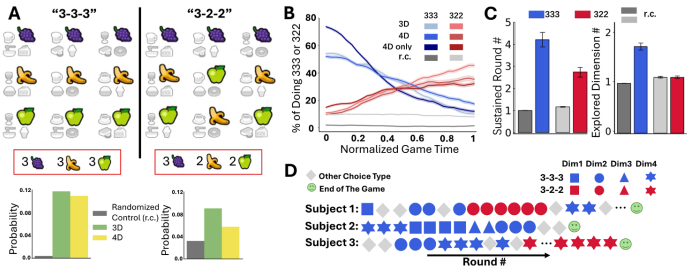

Figure 2: **A)** Illustration for 3-3-3 and 3-2-2 choice patterns and probability of occurrence. **B)** Probability of adopting each choice pattern for different groups on time course. **C)** Persistence and diversity of each choice pattern. **D)** Examples of three subjects' choice sequences in games.

Humans exhibit two stereotyped choice patterns during the game, both occurring at frequencies significantly higher than chance. One pattern, referred to as the "333" choice pattern, involves placing one food three times in each row within a single dimension (**Figure 2A**). Another pattern, termed "322", consists of placing one food three times in the first row and two other foods within the same dimension twice each in second and third rows, respectively (**Figure 2A**). Participants applied these patterns across different dimensions and often switch between dimensions in a sequential manner (**Figure 2D**). Early in the game, the 333 pattern was used more frequently, whereas the 322 pattern became more prevalent in later stages (**Figure 2B**). Both patterns persisted significantly above chance throughout of the game. The 333 pattern exhibited a high degree of diversity, while the 322 pattern showed a certain degree of repetition (**Figure 2C**).

Persistent choice patterns reflect exploration and exploitation along task dimensions

We quantified the preference level of each food in a choice pattern (total of 9 foods in 3D games and 12 in 4D games). To estimate the optimal choice vector for the current round, we computed the correlation between participants' past choices and their corresponding score feedback. The Exploration Index (EI) was then defined as the distance between the actual current choice vector and the predicted optimal vector (**Figure 3A**). Figure 3B illustrates one participant's choice patterns across four dimensions, along with the corresponding EI values. Notably, the EI was higher when the participant adopted the 333 choice pattern and lower when the 322 pattern was used. A significant difference in the EI distributions between the two patterns suggests that the 333 pattern is more strongly associated with exploration, whereas the 322 pattern is less associated with exploration (**Figure 3C**).

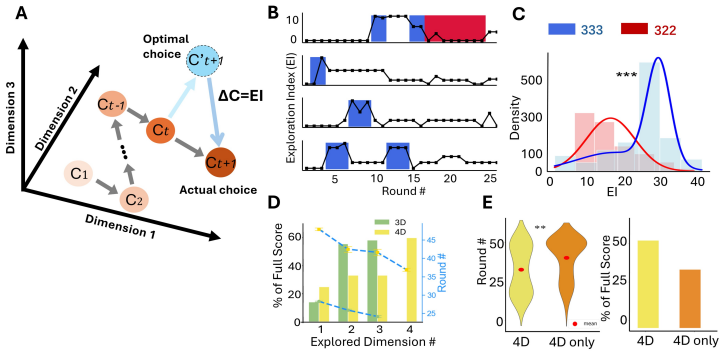

Figure 3: **A)** Illustration of Exploration Index metric. **B/C)** Exploration Index for 333 and 322 choice pattern (case analysis/statistic result). **D)** Efficiency of dimension-guided exploration. **E)** Game performance for naïve and experienced 4D group.

Dimension-guided exploration promotes efficient learning and transfer of strategies

Participants who employed the 333 strategy to explore a greater number of dimensions completed the game more efficiently — in fewer rounds and with higher scores (**Figure 3D**). Those who had previously played the 3D game began using the 333 strategy earlier in the 4D game, indicating a transfer of strategy (**Figure 2B**). A comparison between the experienced and naïve groups in the 4D game revealed that this strategy transfer facilitated more effective learning, as reflected in higher fraction of full-score participants and fewer rounds required to complete the game (**Figure 3E**).

Conclusion

Our results suggest that humans develop a dimension-guided exploration strategy in complex task settings, which is both efficient and transferable. We also observed that this strategy is associated with human category concepts—an aspect that is beyond the scope of the current report. Moving forward, we aim to further investigate how this strategy emerges and to model the underlying cognitive processes involved.

References

Akaishi, R., Kolling, N., Brown, J. W., & Rushworth, M. (2016). Neural mechanisms of credit assignment in a multicue environment. *Journal of Neuroscience*, *36*(4), 1096–1112.

Ballard, I., Miller, E. M., Piantadosi, S. T., Goodman, N. D., & McClure, S. M. (2018). Beyond reward prediction errors: Human striatum updates rule values during learning. *Cerebral Cortex*, *28*(11), 3965–3975.

Farashahi, S., Rowe, K., Aslami, Z., Lee, D., & Soltani, A. (2017, 11). Feature-based learning improves adaptability without compromising precision. *Nature Communications*, *8*. doi: 10.1038/s41467-017-01874-w

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Krause, A., & Ong, C. (2011). Contextual gaussian process bandit optimization. *Advances in neural information processing systems*, *24*.

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*(2), 451–463.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic bulletin & review*, *22*(5), 1193–1215.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208.

Marković, D., Gläscher, J., Bossaerts, P., O'Doherty, J., & Kiebel, S. J. (2015). Modeling the evolution of beliefs using an attentional focus mechanism. *PLoS computational biology*, *11*(10), e1004558.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157.

Parpart, P., Schulz, E., Speekenbrink, M., & Love, B. C. (2017). Active learning reveals underlying decision strategies. *BioRxiv*, 239558.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2018). Putting bandits into context: How function learning supports decision making. *Journal of experimental psychology: learning, memory, and cognition*, *44*(6), 927.

Wang, S., & Rehder, B. (2017). Multi-attribute decision-making is best characterized by an attribute-wise reinforcement learning model. *BioRxiv*, 234732.

Wilson, R. C., & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in human neuroscience*, *5*, 189.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature human behaviour*, *2*(12), 915–924.

Wunderlich, K., Beierholm, U. R., Bossaerts, P., & O'Doherty, J. P. (2011). The human prefrontal cortex mediates integration of potential causes behind observed outcomes. *Journal of neurophysiology*, *106*(3), 1558–1569.