Prospective planning and retrospective learning in a large-scale combinatorial game

Ionatan Kuperwajs (ik1125@nyu.edu)
Center for Neural Science, New York University
New York, NY, United States

Bas van Opheusden (basvanopheusden@nyu.edu)
Department of Computer Science, Princeton University
Princeton, NJ, United States

Wei Ji Ma (weijima@nyu.edu)
Center for Neural Science and Department of Psychology, New York University
New York, NY, United States

Abstract

What algorithms do people use to make decisions with future consequences in complex environments? In order to investigate the cognitive processes underlying sequential planning, we collected large-scale behavioral data in a challenging variant of tic-tac-toe. This task is at an intermediate level of complexity, providing rich behavior for which modeling is still tractable. We argue that a data set of this nature is necessary for distinguishing theoretical frameworks for integration between prospective and retrospective decision-making, and show preliminary evidence for the existence of both systems in our task. We outline a computational model based on an intuitive value function and decision tree search to demonstrate that people engage in prospective planning. We then explain discrepancies between the model's predictions and observed data in early game choices, finding behavioral patterns consistent with retrospective learning.

Keywords: sequential decision-making; planning; reinforcement learning; behavioral modeling

Introduction

Reinforcement learning (RL) is arguably the most successful theoretical framework available for explaining human sequential decision-making and planning (Sutton & Barto, 2018). A central finding in the human RL literature is that people can select actions by combining information from prospective and retrospective systems. To choose an action in a given state, the prospective system mentally simulates the consequences of possible actions multiple steps into the future, whereas the retrospective system considers the outcome of actions taken in the same or similar states in past experience. These dual systems have been discussed under various names and implementations, such as deliberative and habitual (Dolan & Dayan, 2013), goal-directed and Pavlovian (Huys et al., 2012), and model-based and model-free (Daw et al., 2005). In general, the prospective system is slow and computationally expensive, but can determine high-value actions from any state, including ones that the agent has never previously encountered. On the other hand, the retrospective system is fast but needs previous experience to inform its policy.

One outstanding question is how people combine information from these systems or decide whether prospective planning is worthwhile in terms of time and computational resources. This meta-level decision may be based on uncertainty estimates provided by both systems (Daw et al., 2005), the historical accuracy of their predictions (Kool et al., 2017), or estimation of the value of information gained by planning (Callaway et al., 2018; Sezener et al., 2019). A related problem is how these systems can benefit from each other's computations. An appealing framework for integrating prospective planning and retrospective learning is amortization (Dasgupta et al., 2018), in which the agent re-uses simulated experience from the prospective system as additional training data for the retrospective system. Similarly, the retrospective system might influence the prospective system by adapting its internal models and search heuristics.

Here, we establish that people engage in prospective planning by fitting a computational model to their choices in a combinatorial game. We illustrate two behavioral patterns that the model fails to predict, but which are consistent with retrospective learning. While we have not yet developed a complete theoretical framework to fit our data set, we argue that its size and complexity is necessary for understanding the integration of prospective and retrospective RL in a naturalistic setting.

Task

An ideal experimental task for comparing models of prospective and retrospective integration needs to satisfy multiple conditions:

1. The task needs to be novel, so that participants start with uninformed priors to initialize their retrospective system.
2. The task needs to exhibit a large state space so that participants will continually encounter novel states irrespective of experience level, thereby necessitating prospective planning.
3. The task needs to contain a natural division into phases where either prospective or retrospective strategies are likely to be more effective.
4. Participants need to perform the task over long periods of time, since shifts between strategies might require extensive experience.
5. The data set needs to contain many participants, so that some states occur often enough to enable nuanced statistical analyses of people's changing action distributions.

We designed a combinatorial game in which two players alternate placing pieces on a 4-by-9 board, attempting to connect 4-in-a-row. The game has a large state space ($1.18 * 10^{16}$ possible states), and naturally encourages people to adopt retrospective strategies for early-game moves and prospective strategies in the middle and late game (Figure 1A). In the opening, forward search is inefficient, since a decision tree leading to terminal states is necessarily deep and wide. Informative heuristics are also difficult to find, as the empty board contains no patterns. The user always plays first, so opening sequences are likely to repeat across different games, and people can learn an opening policy by trial-and-error. In other words, people are encouraged to develop an "opening book": a tabular representation of state-action mappings which can be updated by model-free RL. By contrast, in the middle and late game, positions are unlikely to ever repeat, but board states tend to contain more patterns and be closer to terminal states, favoring forward planning over retrospective learning.

Additionally, we partnered with Peak, a cognitive exercise company based in London, to implement the game on their mobile platform (https://www.peak.net). We are currently collecting data at a rate of 1.5 million games per month, and here we analyze a subset consisting of approximately 3.2 million games from 430,000 unique users. Users play against an AI agent implementing a version of our computational model, with parameters adapted from fits on previously collected human-vs-human games (van Opheusden et al., 2017).

Computational model

In order to demonstrate that users engage in prospective planning, we developed a computational model which combines tree search with a feature-based value function, stochastic feature dropping, and value-based pruning (van Opheusden et al., 2017).

Value function

The core component of our model is an evaluation function $V(s)$ which assigns values to board states s. We use a weighted linear sum of 5 features: center, connected 2-in-a-row, unconnected 2-in-a-row, 3-in-a-row and 4-in-a-row. The center feature assigns a value to each square corresponding to inverse Euclidean distance from the board center, and sums up the values of all squares occupied by the player's pieces. The other 4 features count how often the associated pattern occurs on the board. We associate weights w_i to these features, and define

$$V(s) = c_{\text{self}} \sum_{i=0}^{4} w_i f_i(s, \text{self}) - c_{\text{opp}} \sum_{i=0}^{4} w_i f_i(s, \text{opponent})$$

where $c_{\text{self}} = C$ and $c_{\text{opp}} = 1$ when it is the player's move, and $c_{\text{self}} = 1$ and $c_{\text{opp}} = C$ when it is the opponent's move. C captures value differences between active and passive features. For example, a three-in-a-row feature signals an immediate win on the player's own move, but not the opponent's.

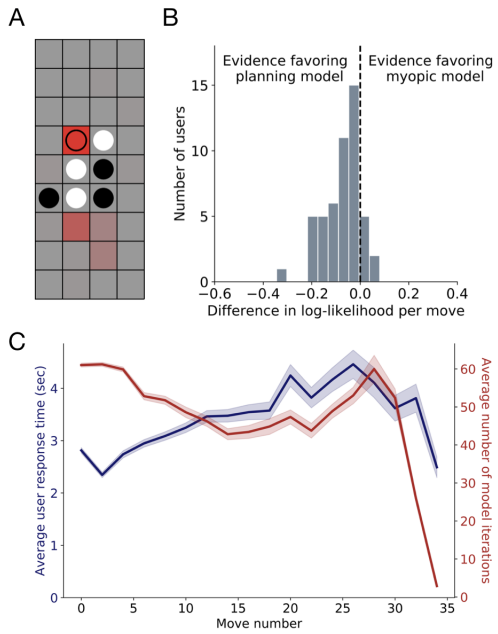

Figure 1: Model performance. **(A)** An example board with model predictions. The red shading indicates the probability distribution of the model's move prediction and the open circle indicates the user's move. **(B)** Histogram of the difference between cross-validated log-likelihoods per move for the planning and myopic models. **(C)** Average user response times (blue) and average number of model iterations (red) taken to make a move throughout gameplay. Axes are scaled by the maximum value in each set of averages for visualization purposes, and shading denotes s.e.m.

Tree search

The evaluation function guides the construction of a decision tree with an iterative best-first search algorithm. Each iteration, the algorithm chooses a board position to explore, evaluates the positions resulting from each legal move, and prunes all moves with value below that of the best move minus a threshold θ. The algorithm has a stopping probability γ, resulting in a geometric distribution over the number of iterations.

Noise

To account for variability in people's choices, we add three sources of noise. Before constructing the decision tree, we randomly drop features at specific locations and orientations, which are omitted during the calculation of $V(s)$. During tree search, we add Gaussian noise to $V(s)$ at each node. Finally, we include a lapse rate λ.

Model fitting

When fitting the computational model to behavioral data, we infer parameters for individual users with maximum-likelihood estimation. The model has 10 parameters: the 5 feature

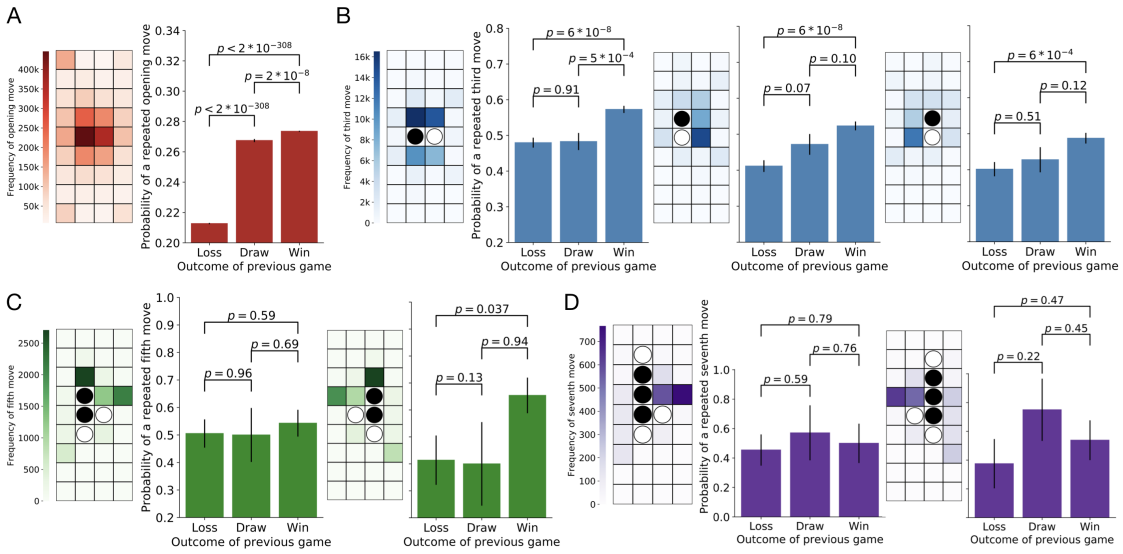

Figure 2: The probability that users repeat a move directly after a loss, draw, or win given different board states and the distribution of the selected moves. Error bars denote s.e.m. The user piece is in black while the AI piece is in white. **(A)** The opening move. **(B)** The third move following the most frequent 2-piece board states. **(C)** The fifth move following the most frequent 4-piece board states. **(D)** The seventh move following the most frequent 6-piece board states.

weights, the active-passive scaling constant C, the pruning threshold θ, stopping probability γ, feature drop rate δ, and the lapse rate λ. We estimate the log probability of a user's move in a given board position with inverse binomial sampling, and optimize the log-likelihood function with multilevel coordinate search. We account for potential overfitting by reporting 5-fold cross-validated log-likelihoods, with the same testing-training splits for model comparison.

Results

Evidence for prospective planning

The average accuracy of the computational model's predictions on the hold-out data is $23.5 \pm 0.8\%$, which is much better than chance ($5 \pm 0.1\%$). To test whether the tree search component is necessary to fit human choices, we compared the model's log-likelihood per move with that of a myopic model. In the myopic model, we fix γ to 1, which implies that the tree search terminates after a single iteration. Because model fitting and comparison is computationally taxing, we ran this analysis on 50 pseudo-randomly selected users. The cross-validated log-likelihood per move of the computational model is significantly higher across users than that of the myopic model ($t = 6.69$, $p = 2 * 10^{-8}$, Figure 1B), demonstrating that tree search is indeed necessary to predict people's moves.

Discrepancies between data and the model

One major difference between the model and our observed data is predicted response times. Previously, we found that people's response times correlate on individual trials with the number of model iterations (van Opheusden et al., 2017). However, their average trend over the course of a game differs considerably (Figure 1C). Early in gameplay, the model predicts that people search larger decision trees and thus have longer response times, but the data shows the opposite. Therefore, it is likely that in situations where the board is fairly empty and no player can immediately win the game, there is a faster retrospective process that takes place before prospective planning begins. In the middle and late game, response time trends roughly follow model predictions.

Evidence for retrospective learning

The size of our data set allowed us to uncover clear evidence for retrospective learning in early-game positions. We found that users were significantly more likely to repeat their opening moves following wins rather than losses, and that these moves were primarily distributed in the center or corners of the board (Figure 2A). This effect continued on the third move, where users most often elected to play in the center positions closest to the two pieces already on the board (Figure 2B). On the fifth and seventh moves, however, the proportion of move repetitions based on game outcome decreased, and varied by specific board position despite consistent move selections (Figure 2C,D). These population-wide trends suggest that people make decisions partially based on whether or not an opening strategy was successful in previous games in their first two moves, and then begin to utilize alternative strategies in subsequent moves when board positions are more likely to be unique.

Next, we show that response times in early stages of a game also follow patterns predicted by retrospective learning. This was similarly mediated by previous game outcome: user

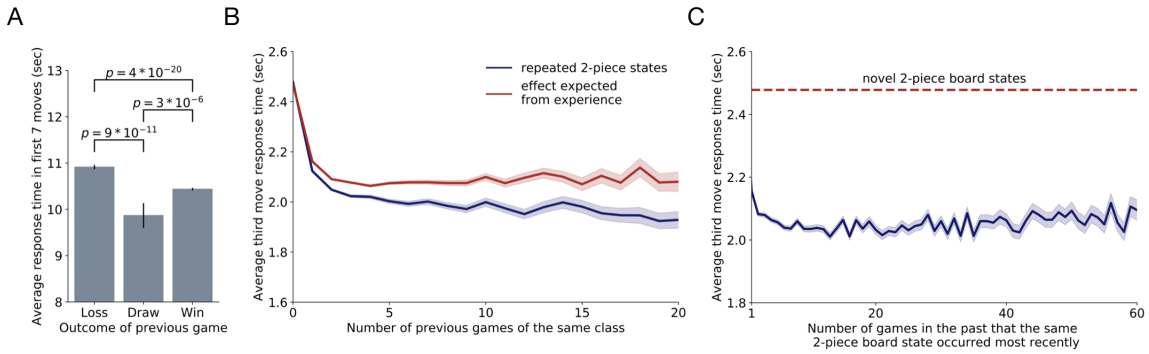

Figure 3: User response times. Error bars and shading denote s.e.m. **(A)** Average response times across the first 7 moves of a game directly following a loss, draw, or win. **(B)** Average response times on the third move as a function of repeated 2-piece board states (blue), and the average response time of 1000 randomly sampled users that had previously played the same number of games (red). **(C)** Average third move response times as a function of the number of games in the past that the same 2-piece board state occurred (blue) compared to novel 2-piece board states (red).

response times across the first 7 moves were, on average, longer after losses rather than wins (Figure 3A). Furthermore, third move response times decreased significantly when users encountered repeated 2-piece board states (Figure 3B). This could be a confounded result, since on average users play faster after playing multiple games regardless of which states occurred. Therefore, we ran a control in which we sampled the average response times of other users that had played the same number of games, explaining some of the effect but not all. Finally, we verified that the effect was not solely due to recent memory of encountered states. We averaged third move response times based on the number of games in the past that the same 2-piece board state occurred, and found that response times were consistent regardless of how long ago a given state had been seen (Figure 3C). These response times were also drastically lower than for novel 2-piece board states.

Discussion

In this article, we analyze human behavioral data in a two-player combinatorial game, and find strong evidence for both prospective planning and retrospective learning. We demonstrate that a computational model based on a forward search algorithm fits human choices well in the middle and late game, but not the early game. However, we find that people's early-game moves as well as their response times are affected by the outcome of previous games in which they encountered the same board positions. These results demonstrate that people learn from past experience and are consistent with many retrospective learning algorithms, ranging from simple win-stay-lose-shift to sophisticated policy gradient methods. Our findings suggest that people strategically integrate information from a prospective and a retrospective system, and that a data set of this nature is essential for differentiating between existing theoretical frameworks of prospective and retrospective integration.

Acknowledgments

This work was supported by grant IIS-1344256 from the National Science Foundation and grant 1R01MH118925-01 from the National Institutes of Mental Health.

References

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P., & Griffiths, T. (2018). A resource-rational analysis of human planning. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Dasgupta, I., Schulz, E., Goodman, N., & Gershman, S. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, *178*, 67–81.

Daw, N., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704–1711.

Dolan, R., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*, 312–325.

Huys, Q., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. (2012). Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLOS Computational Biology*.

Kool, W., Gershman, S., & Cushman, F. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*, 1321–1333.

Sezener, C., Dezfouli, A., & Keramati, M. (2019). Optimizing the depth and direction of prospective planning using information values. *PLOS Computational Biology*.

Sutton, R., & Barto, A. (2018). *Reinforcement learning: An introduction, 2nd edition*. MIT Press.

van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. (2017). A computational model for decision tree search. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.