

Good and consequential counterfactual outcomes are prioritized during learning

Kate Nussenbaum (katenuss@princeton.edu) & Nathaniel Daw (ndaw@princeton.edu)

Princeton Neuroscience Institute, Princeton, NJ, 08544, USA

Abstract

People can learn from actions not taken by leveraging mental models to imagine their potential consequences. However, for any given choice, the number of possible, alternative actions often exceeds the brain's capacity for simulation. Here, we develop a new task to measure behaviorally whether people selectively prioritize the counterfactual updates that are most likely to improve their future decisions. Our initial results (N = 69) indicate that people most strongly consider high-magnitude alternatives as well as those that are better than the option they selected, suggesting that people do indeed consider alternative possibilities strategically.

Keywords: reinforcement learning; model-based learning; mental simulation

Introduction

People learn from direct experience, by taking actions and observing their positive and negative outcomes. They also learn – and perhaps even more so – from indirect experience, by harnessing mental models to simulate the consequences of the myriad actions they *could have* taken (Daw, Niv, & Dayan, 2005). However, model-based learning poses a potential problem: In most situations, there is a near-infinite number of possible actions to consider, and people do not have the time or mental resources to think about all possible alternatives. How do people select possibilities to think about? Do they judiciously manage their thoughts, bringing to mind and updating their beliefs in such a way as to most effectively improve their future decisions?

Theoretical work has argued that simulation is valuable to the extent that it is likely to improve future decisions (“gain”), and particular options should be prioritized for simulation on this basis (Mattar & Daw, 2018). Prior work has primarily investigated this question using sequential decision-making tasks, in which people can leverage structured knowledge to propagate experienced value to more distal parts of physical or abstract state spaces (Liu, Mattar, Behrens, Daw, & Dolan, 2021; Schuck & Niv, 2019; Ophesusden et al., 2023).

Although sequential decision-making tasks have yielded important insights into learning and planning, they nonetheless have several downsides for investigating how people manage their thoughts. First, in these tasks, it is often difficult to ascertain from behavior alone which specific trajectories people simulate. Many studies of simulation or replay have thus relied upon neuroimaging methods to characterize patterns of thought (Liu, Dolan, Kurth-Nelson, & Behrens, 2019; Liu et al., 2021; Schuck & Niv, 2019), which are costly and difficult to administer to large and diverse samples of participants.

Here, we sought to develop a simpler task to behaviorally characterize the alternative actions people bring to mind and learn about. In doing so, we built upon a different line of past studies, using structured, multiplayer games, which can also provide insight into how people engage in counterfactual, model-based updating over a set of candidate moves. One previous study showed that people preferentially update “upward counterfactual” options (Hunter, Meer, Gillan, Hsu, & Daw, 2022), i.e. those that would have been better than the action actually chosen, in line with rational prioritization (Mattar & Daw, 2018) because these are the options associated with actionable policy improvements. However, the specific game payoff matrices used in that study limited the ability to study finer grained effects, including particularly the idea that not just the direction of update, but also the magnitude, should drive prioritization (Moore & Atkeson, 1993). Our goal was to develop a fun and engaging task that could more finely index individuals’ prioritization *strategy*, and also separate it from their *capacity* for bringing alternative possibilities to mind.

Methods

Young adult participants (N = 69, ages 18 - 30 years, recruited from Prolific) completed the ‘Battle Cards’ task online. Briefly, on every turn, both the participant and their computerized opponent selected one of five element cards to play (Fig. 1A). Each card would win or lose from 2 to -2 points against each other card (Fig. 1A). Participants were told that each of their opponents had different preferences for the cards they liked to play, and that their opponents had all pre-determined their decks ahead of time and would not adjust their strategy in response to their choices. Thus, while no card was inherently better or worse than any other, participants could earn more points by learning their opponents’ preferences, and selecting those most likely to win against those that their opponents frequently played. Critically, on every trial, participants could learn through direct experience by updating their belief about the value of the card they played based on the points they won or lost. However, because the task had a known, deterministic payoff structure, participants could also learn through indirect experience, by bringing to mind each of the four alternative cards they could have played and imagining its associated counterfactual outcome.

Prior to engaging in ‘real’ gameplay, participants went through extensive instructions that explained the game mechanics as well as the relations among the cards, which were designed to be as intuitive as possible (e.g., Water defeats Fire, Fire defeats Grass, etc.). In addition, they completed two tutorials to ensure that they well learned the task’s payoff matrix. In the first tutorial (100 trials), participants were shown the

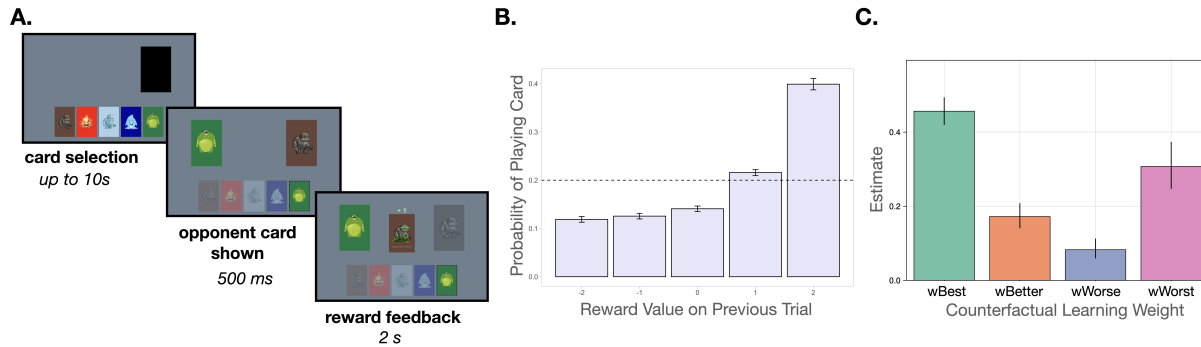


Figure 1: A) Participants played eight games of 20 trials each, in which they selected among five cards to try to defeat the computer’s card. B) Participants were more likely to play cards that would have earned more points on the preceding trial. C) Participants more heavily weighted high-magnitude and better outcomes when learning about counterfactual alternatives.

card their opponent played and then had to select which card to play. In the second tutorial (100 trials), participants were shown two cards and asked to select the one that would win the battle. In both tutorials, participants received explicit feedback on their choices. Finally, participants played eight blocks of ‘real’ games in which they faced eight different opponents for 20 trials each.

Results

Evidence for reward-seeking behavior

First, we confirmed that participants learned the task’s payoff matrix throughout the two tutorials. In the last 10 trials of the first tutorial, participants earned on average 1.81 points (SE = .04 points) points per trial, close to the maximum of 2 points per trial, and well above chance-level performance (0 points per trial). In the second tutorial, participants selected the better card on 9.6 (SE = .01) out of the last 10 trials, indicating that they effectively learned the game structure.

We then analyzed whether participants learned to select cards to defeat each opponent during real game-play. Participants earned on average 6.26 points per game, performing significantly above chance-level (0 points per game), $t(68) = 9.88, p < .001$. In addition, participants were more likely to play cards that *would have* earned higher rewards on the prior trial (Fig. 1B), indicating that they chose cards in response to their opponents’ selections.

Prioritization of better and high-magnitude alternatives

Next, we asked *how* participants learned to play rewarding cards – to what extent did they a.) rely on a mental model of the game to learn about counterfactual options and b.) prioritize *specific* counterfactual options for updating? To answer these questions, we fit our data with a reinforcement-learning model. On every trial, the model assumed that participants updated their belief about the option they selected with a model-free learning rate, as well as about the four counterfactual options. Critically, the counterfactual learning rate for each alternative was determined by multiplying the model-free

learning rate by one of four weighting parameters: one for the *best* alternative (i.e., the option that would have earned +2), one for *better* alternatives (i.e., other options that would have earned more points than the player actually earned), one for *worse* alternatives (i.e., other options that would have earned less points than the player earned), and one for the *worst* alternative (i.e., the option that would have earned -2).

We fit the model hierarchically and examined group-level parameter estimates for the weights. All four weights were above 0 (Fig. 1C), indicating that participants harnessed their knowledge of the task’s payoff matrix to learn in the absence of direct experience. Further, the pattern of weights revealed that participants did not just consider alternatives at random. Replicating (Hunter et al., 2022), alternatives better than the chosen one were more strongly updated than worse ones, and in addition, among both worse and better action sets, higher-magnitude alternatives (i.e., the best and worst options) were updated to a greater extent.

Discussion

Our initial data indicate that our novel card game task can effectively index complex patterns of model-based thought in the absence of neural measures. In addition, our modeling results suggest that participants may have engaged in strategic simulation, selectively considering those options that were likely to improve their future decisions the most. Intuitively, agents can maximize future reward by learning about the options they *should have* chosen (and therefore, should choose in the future) as well as by learning about the options that will affect their reward gain the most. Indeed, preliminary simulations of a cognitively implausible, “gain”-based agent – which computes the extent to which performing each counterfactual update will improve subsequent decisions and then selects updates accordingly – revealed a qualitatively similar pattern of counterfactual learning rates. In future work, we will further examine the extent to which participants’ learning aligns with the predictions of “rational” models, as well as how dimensions of individual variance (age, psychopathology) relate to simulation capacity vs. strategy.

Acknowledgments

We gratefully acknowledge our funders: the CV Starr Foundation (Fellowship to K.N.) and the NIMH (grant MH135587 to N.D., part of the CRCNS program).

References

- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.
- Hunter, L. E., Meer, E. A., Gillan, C. M., Hsu, M., & Daw, N. D. (2022). Increased and biased deliberation in social anxiety. *Nature Human Behaviour*, *6*(1), 146–154. doi: 10.1038/s41562-021-01180-y
- Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. (2019). Human Replay Spontaneously Reorganizes Experience. *Cell*, *178*(3), 640–652.e14. doi: 10.1016/j.cell.2019.06.012
- Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D., & Dolan, R. J. (2021). Experience replay is associated with efficient nonlocal learning. *Science*, *372*, eabf1357. doi: 10.1126/science.abf1357
- Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*(11), 1609–1617. doi: 10.1101/225664
- Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning*, *13*, 103–130.
- Opheusden, B. v., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, *618*(7967), 1000–1005. doi: 10.1038/s41586-023-06124-2
- Schuck, N. W., & Niv, Y. (2019). Sequential replay of non-spatial task states in the human hippocampus. *Science*, *364*(6447). doi: 10.1126/science.aaw5181