Some and Done? Temporally extended decisions with very few rollouts

Sixing Chen (sixing.chen@nyu.edu)
Department of Psychology, New York University

Kristopher T. Jensen (kris.jensen@ucl.ac.uk)
Sainsbury Wellcome Centre, University College London

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

Abstract

It has been suggested that humans mentally simulate the outcomes of their actions when making decisions. However, this process can be challenging in real-world decision-making, which typically involves temporally extended decision trees with numerous potential outcomes. Here, we demonstrate with a computational model that temporally extended decisions can be achieved with just a few forward simulations, formalized as rollouts. We also show that, under resource constraints, performing many partial (shallow) rollouts can yield more favorable outcomes than performing fewer full (deep) rollouts. Additionally, our model captures behaviors traditionally attributed to pruning or satisficing strategies without the need for explicit heuristics, providing an alternative explanation for these phenomena. Finally, we show that the dynamics of value estimation over successive rollouts closely resemble evidence accumulation models. Our framework offers a plausible mechanism for temporally extended decision-making and provides insights into the neural underpinnings underlying this process.

Keywords: decision-making; mental simulation; reinforcement learning; planning

Introduction

Previous research has suggested that humans use mental simulation for evaluating the outcomes of their actions when making decisions (Battaglia, Hamrick, & Tenenbaum, 2013). In simple decision problems, in particular, studies suggest that humans rely on very few samples of simulated outcomes to inform their decisions (Erev & Roth, 2014; Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Erev, 2009). While this may appear implausible at first, an influential theory suggests that decisions based on very few samples are not only possible, but the optimal strategy under the assumption that samples are costly (Vul, Goodman, Griffiths, & Tenenbaum, 2014). Yet, despite theoretical and empirical evidence that few samples are needed in simple decision problems, it is not clear whether the same conclusion holds in more general scenarios, such as in temporally extended decision-making.

In temporally extended decision-making scenarios, mental simulation involves navigating through complex decision trees that have numerous potential outcomes. Consider a temporally extended decision problem with a branching factor of b at every time step. Simulating n steps ahead involves evaluating b^n possible paths, which quickly becomes intractable as b or n grows. Given the constraints of time and computational power in human cognition, such extensive forward simulation seems at odds with our ability to make rapid and accurate decisions in these extended scenarios. Although it is believed that humans can leverage a wide variety of strategies to reduce the computational load (Callaway et al., 2022; Huys et al., 2015), the complexity of the problem could still mean that effective forward simulation would demand an impractical level of resources.

In this work, we extend the ideas of Vul et al. (2014) to temporally extended decision trees and investigate whether good decisions can still be achieved with very few forward simulations. We formalize a forward simulation as a rollout – an iterative sampling procedure, where each sampled state is always a successor (child) of the previously sampled state, and which terminates when the most distant state is sampled. We consider the setting in which the agent selects a single action that results in a series of stochastic transitions that the agent has no control over. In this setting, we show that the optimal approach in temporally correlated environments is to decide on the basis of a fairly small number of rollouts. We also find that shallower rollouts are favored when computational resources are limited, and that deeper rollouts are favored when resources are abundant. Additionally, our model predicts a tendency to prefer paths with low variance in rewards, offering an alternative explanation for behaviors traditionally attributed to pruning or satisficing strategies. The model also has dynamics that closely resemble evidence accumulation models, giving clear predictions of reaction time patterns in temporally extended decision trees. Finally, we show that our conclusions hold in a wide range of environments, and investigate when more rollouts are needed. In particular, we show that reducing the correlations in cumulative rewards between different paths leads to the need for more rollouts.

Model

Task environment

We assume an environment described by a decision tree with a branching factor of b and a depth of d. The agent must choose an action at the root node. Each node i is associated with a Gaussian-distributed reward R_i:

$$R_i \sim \mathcal{N}(\mu_i, \sigma_R^2), \tag{1}$$

where the mean of the Gaussian distribution is randomly initialized:

$$\mu_i \sim \mathcal{N}(0, \sigma_\mu^2). \tag{2}$$

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

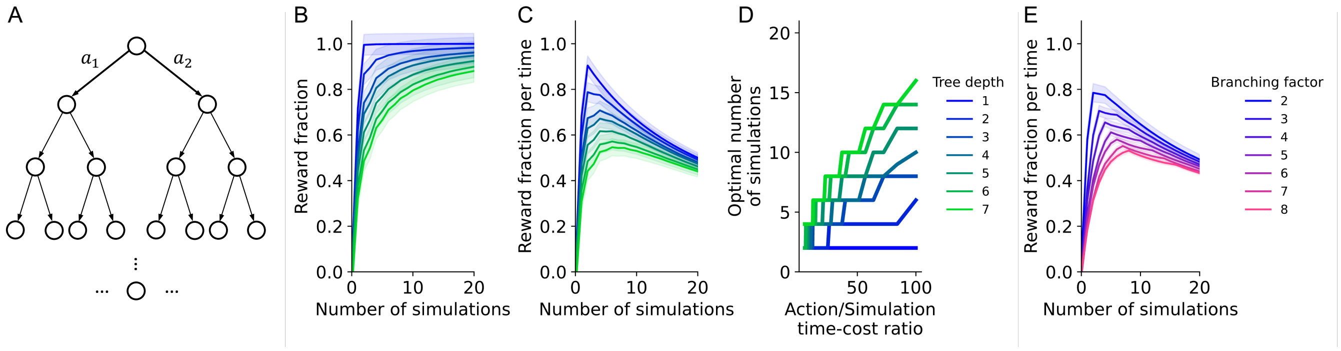

Figure 1: Optimal decisions based on very few rollouts. (A) Task illustration with a branching factor of $b = 2$. The agent chooses an action at the root node to deterministically transition to the corresponding first-layer node and then goes through random transitions in subsequent nodes. (B) Fraction of maximal reward achieved as a function of the number of rollouts performed before taking an action. (C) Fraction of maximal reward achieved divided by the time needed to execute all the rollouts and one action, as a function of the number of rollouts performed before taking the action. (D) Optimal number of simulations that yield the highest reward fraction per time as a function of action/simulation time-cost ratio. (E) The same as (C) but under different branching factors. Here we illustrate the results with a tree depth of $d = 2$.

The cumulative reward $G_i = \sum_{i' \in \text{ancestors}(i)} R_{i'}$ obtained from selecting a root-node action and traversing the tree to a leaf node i is therefore distributed according to:

$$G_i \sim \mathcal{N}(m_i, \sigma_G^2), \tag{3}$$

where

$$m_i = \sum_{i' \in \text{ancestors}(i)} \mu_{i'}, \tag{4}$$

and

$$\sigma_G^2 = d\sigma_R^2. \tag{5}$$

This leads to a reward correlation structure in the environment, where leaf nodes closer to each other in the decision tree tend to have similar cumulative rewards.

Agent

We assume that an agent needs to select an action at the root node that maximizes the expected cumulative reward (Figure 1A). While the root-node policy is thus controlled by the agent, we assume that all subsequent transition probabilities are fixed. The agent is equipped with a world model consisting of ground truth reward and transition probabilities and can draw samples from the model. Each sample is a forward rollout from the root node to one of the leaf nodes, which provides a noisy estimate of the cumulative reward for an action. Specifically, the agent chooses a root-node action a for a rollout and gets a sample s from the cumulative reward distribution $p(G|a)$, where

$$p(G|a) = \sum_i p(G|i)p(i|a). \tag{6}$$

Here, $p(G|a)$ is the cumulative reward distribution conditioned on choosing the root-node action a for a rollout, marginalized across all leaf nodes that the rollout could visit.

Before selecting an action, the agent draws a finite number of samples S from the world model. We assume that the agent draws samples with a rollout policy that sequentially samples each action. This sequential rollout policy assumes no systematic selection of root-node actions and samples all actions as evenly as possible. In the simplest case, we also assume that the transition distribution is uniform at all nodes beyond the root node. This uniform transition distribution implies that all leaf nodes are equally likely to be visited by a rollout for a given root-node action.

After sampling the rollouts, the agent takes the action with the highest estimated cumulative reward based on the sampled rollouts:

$$a^* = \underset{a}{\text{argmax}} \left[\frac{1}{|S_a|} \sum_{s \in S_a} g_s \right], \tag{7}$$

where S_a is the subset of rollouts from the root-node action a and g_s is the cumulative reward of the sampled rollout s.

Results

Optimal decisions based on very few rollouts

We first investigated how many samples an agent should draw to optimize its decisions in a temporally extended setting by simulating our model with a branching factor of $b = 2$. We computed the expected reward fraction under the resulting policy, defined as the fraction of the maximal available reward that was achieved by the agent in a given environment. When repeating this analysis across tree depths, the expected reward increased with the number of simulations, reflecting the fact that each sample reduces uncertainty about the optimal policy (Figure 1B). Thus, as long as sampling is "free", an agent should sample infinitely many rollouts before making a decision.

We next considered a scenario where rollouts take time, resulting in an opportunity cost. Suppose the agent is given a fixed amount of time to freely conduct sampling and take actions. The more samples the agent draws before making a choice, the fewer actions it can take in a fixed amount of time. In this case, if rollouts are as slow as physical actions, no samples should be drawn, since the information gained by sampling and acting is identical, yet only the latter results in real rewards. However, if a sample can be obtained in less time, the increased rewards expected from sampling can offset the time spent not acting. To capture this idea, we defined ‘expected reward fraction per time’ as the expected reward fraction divided by the time to execute all the rollouts and one action. For a relatively modest simulation cost of 0.05 (i.e., one action takes as long as 20 simulations), we found that the expected reward fraction per time peaked at a relatively small number of simulations (Figure 1C). Indeed, across a range of tree depths and simulation costs, the optimal number of simulations was typically very small, surpassing 10 only when simulations were cheap and trees were deep (Figure 1D). The same pattern of results was also found under different branching factors. (Figure 1E). These results are similar to those reported previously for bandit problems (Vul et al., 2014) and suggest that the optimal approach in the temporally extended decision-making setting is similarly to take action on the basis of a fairly small number of rollouts.

Optimal simulation depth

In the preceding analyses, each sample was a rollout terminating at a leaf node. Yet, due to the stochasticity of the transitions, shallower nodes can be more informative than deeper nodes. This is because a specific node is less likely to be visited the more distant it is from the action. Thus, if we assume that the agent is capable of evaluating the cumulative reward to an intermediate node and has finite computational resources, it may be beneficial to perform partial rollouts in temporally extended decision-making settings.

To investigate this, we assumed a finite computational budget specifying the number of node expansions an agent could simulate before taking an action and considered rollouts of different depths. We found that the agent achieved higher average rewards as the computational budget increased, because more individual rollouts were sampled (Figure 2A). Importantly, however, the optimal rollout depth depended on the computational budget, consistent with our hypothesis. For small budgets, shallow rollouts were typically superior, since they enabled more samples from the earlier, more relevant nodes. However, as the computational budget increased, the additional shallow rollouts provided increasingly less information about earlier nodes while providing no information about deeper nodes. For large computational budgets, therefore, deeper rollouts were superior, providing sufficient information about all nodes of the tree (Figure 2A). This result holds across trees with different branching factors (Figure 2B). In trees with higher branching factors, simulating all early nodes required a larger computational budget, leading to smaller optimal rollout depths (Figure 2B). In summary, the optimal strategy for a capacity-limited agent is to perform shallow rollouts for small computational budgets and deeper rollouts when it has enough resources to “think”.

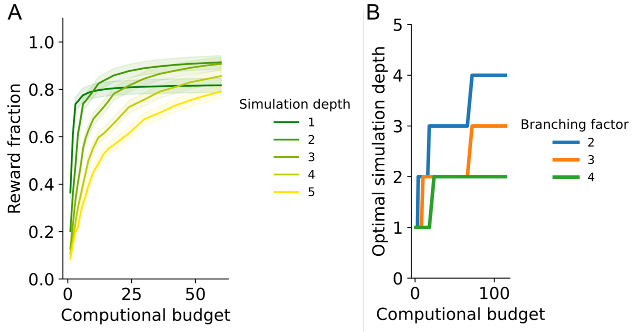

Figure 2: Optimal simulation depth. (A) Fraction of maximal reward achieved as a function of computational budget (the number of node expansions before taking an action) under different simulation depths. Here we illustrate the results with a branching factor of $b = 3$. (B) Optimal simulation depth that yields the highest reward fraction as a function of computational budget for different branching factors.

Pruning and satisficing

We then considered the case where some actions lead to high-variance paths while others lead to low-variance paths. Previous studies suggest two heuristics in human planning in such settings: pruning and satisficing. Specifically, humans stop considering sub-trees when they encounter a large loss (pruning; Huys et al., 2012, 2015), and they commit to a path when they encounter a large reward (satisficing; Simon, 1955). We asked whether our model can provide alternative explanations for previous results on pruning and satisficing without building in assumptions of the corresponding heuristics.

We first considered the case of pruning, where there is an early punishment for one action followed by an even larger reward some time in the distant future. Critically, we now assume that the environment transitions are deterministic and the agent can control its policy at every node. In this case, the agent should take the high-variance action in order to maximize its cumulative reward, provided that subsequent transitions are deterministic. However, if there is uncertainty in the internal world model, the high-variance action may be disfavored. This is because the uncertainty can cause the agent to overestimate the possibility of not getting the later payoff after the initial large punishment. To investigate this phenomenon, we let one action lead to a large punishment of -10 at a depth of $n = 1$ and to a large positive reward of $+15$ in one of the downstream leaf nodes for the action. Additionally, instead of having uniform transition probabilities in the internal model, we set the transition probabilities of moving towards the large reward leaf node to $1 - \varepsilon$ at each node, where ε corresponds to the uncertainty in the world model. This setting

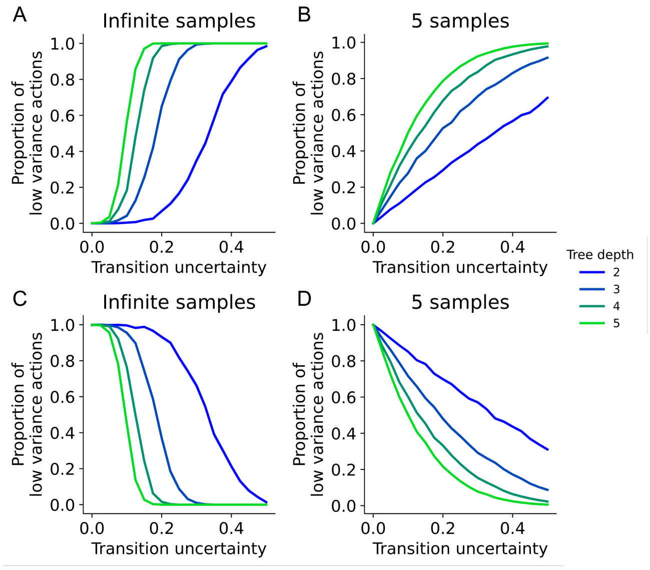

Figure 3: Pruning and satisficing. Transition uncertainty corresponds to the parameter ε that characterizes the probability of failing to transition towards the terminal leaf node with a large reward or punishment at each transition. (A-B) Environment corresponding to the pruning strategy, where there is an early punishment for one action followed by a larger reward later in the tree. Increasing uncertainty leads to favoring the low-variance action. (C-D) Environment corresponding to the satisficing strategy, where there is an early reward for one action followed by a larger punishment later in the tree. Increasing uncertainty leads to favoring the high-variance action. The high-variance action is optimal in A-B, and the low-variance action is optimal in C-D.

can be interpreted as the agent intending to move towards the rewarded leaf node, but the noisy internal model predicts that there is a probability ε of each intended transition failing.

As predicted, the agent favored the high-variance action for small uncertainties. However, as the uncertainty increased, which decreased the estimated probability of being able to visit the rewarded leaf node, the low-variance action became increasingly favored (Figure 3AB). The agent also tended to favor the low-variance action in deeper trees, since the agent had a higher estimated cumulative probability of failing to move towards the rewarded leaf node. These observations hold both when considering the action taken in the limit of infinite samples (Figure 3A) and also when we consider the action taken on the basis of a finite number of samples (Figure 3B). The key difference between these two settings is that the transition between the two actions changes more gradually in the finite-sample situation. This is because, in the finite-sample situation, the estimated probability of transitioning towards the rewarded node is more noisy. These results provide an alternative explanation for previous findings that humans prune decision trees when they encounter large losses during planning (Huys et al., 2012, 2015).

Similarly, we also considered the case for satisficing, where there is an early reward of $+10$ for one of the actions at a depth of $n = 1$, followed by a larger punishment -15 in one of the downstream leaf nodes of the action. In this case, we assume the punishment to be inevitable, but the uncertain world model of the agent suggests a probability ε of escaping this path at each non-root node. We observed symmetric patterns, where the agent favored the risky action for large uncertainty, and the alternative for small uncertainty (Figure 3CD). This is consistent with previous findings that humans commit to sub-optimal options when they encounter large rewards and thus miss optimal options (Simon, 1955).

Importantly, we suggest an alternative underlying mechanism for previous observations of pruning and satisficing. Here, myopic decisions arise not from heuristics, but from a rational evaluation of future outcomes under uncertainty. Our model may not account for all aspects of the pruning and satisficing effects. Specifically, our model is similar to introducing a temporal discount factor, since the finite probability of transitioning towards a non-rewarding node can be captured by such a discount factor. However, Huys et al. (2012) showed that while the pruning effect could be partially explained by a general temporal discount factor, participants also demonstrated additional pruning in response to large negative outcomes, which could only be explained by introducing a second temporal discount factor. We hypothesize that such an effect could potentially be captured by a model where uncertainty increases with planning depth, in contrast to our current framework which has a fixed transition uncertainty throughout the tree.

Relationship with evidence accumulation

We then investigated the connections between our model and evidence accumulation models. Numerous studies in bandit settings suggest that noisy decision-making is based on an evidence accumulation process, where people accumulate samples to reach a decision boundary before making a choice (Bakkour, Zylberberg, Shadlen, & Shohamy, 2018; Gold & Shadlen, 2007; Ratcliff & McKoon, 2008; Zylberberg, Bakkour, Shohamy, & Shadlen, 2024). Moreover, recent studies propose that the same evidence accumulation process also underlies temporally extended decision-making (Solway & Botvinick, 2012). To draw connections between our model and evidence accumulation models, we asked whether our model shows the same accumulation dynamics and whether it can generate new predictions in temporally extended decision trees.

Consistent with previous studies (Gold & Shadlen, 2007; Vul et al., 2014), we considered an environment with a branching factor of $b = 2$ and reframed the decision problem as a hypothesis testing problem. In this case, the agent needs to differentiate between two alternative hypotheses, (i) action 1 is best, and (ii) action 2 is best. The agent performs rollouts and accumulates the resulting information until the cumulative evidence is strong enough to support one hypothesis over the other. We assume that every time step the agent draws a

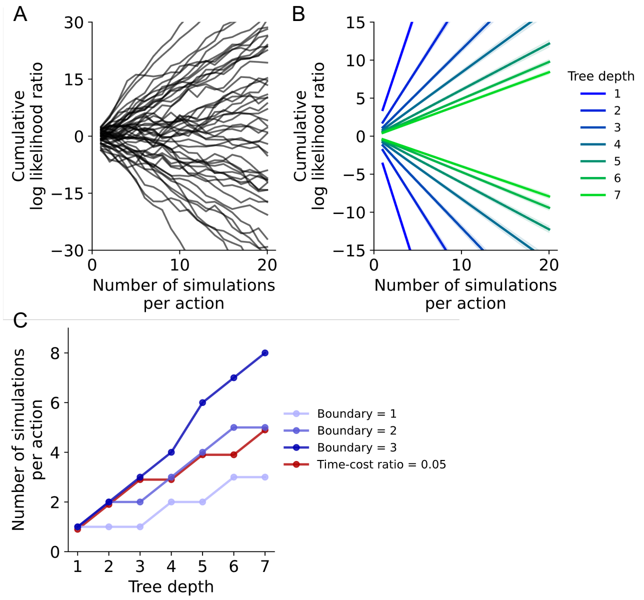

Figure 4: Relationship with evidence accumulation. (A) Evidence accumulation dynamics in individual trials illustrated with a tree depth of $d = 4$. (B) Evidence accumulation dynamics averaged across trials for different tree depths. Curves with positive and negative slopes correspond to the accumulation process under the two hypotheses respectively. (C) Number of simulations per action as a function of tree depth. Blue curves correspond to the predictions of the constant boundary model. The red curve corresponds to the predictions of the model that maximizes reward fraction per time under a time-cost ratio of 0.05.

sample from each action and computes the sample difference $g_1 - g_2$ between the two actions. Then the agent computes the log likelihood ratio of the sample difference under the two alternative hypotheses:

$$\mathcal{L} = \log \frac{p(g_1 - g_2 | \text{option 1 better})}{p(g_1 - g_2 | \text{option 2 better})}. \tag{8}$$

The agent accumulates the log likelihood ratios across samples and makes a decision when the cumulative log likelihood ratio reaches a constant boundary.

We found that the accumulation dynamics in individual trials closely resembled the accumulation dynamics found in previous studies (Bakkour et al., 2018; Gold & Shadlen, 2007; Ratcliff & McKoon, 2008; Zylberberg et al., 2024), where the agent accumulated noisy samples towards one of the two hypotheses (Figure 4A). We then looked into the effect of tree depth on the accumulation process. We found that the accumulation speed was slower for deeper trees, due to the fact that evidence got more ambiguous for deeper trees (Figure 4B). We then considered different decision boundaries and tested how many simulations were needed to reach the boundaries. We found that the number of simulations needed was consistently higher for deeper trees across different decision boundaries, suggesting that reaction times should be longer for deeper trees (Figure 4C). Moreover, we compared the number of simulations predicted by both the constant boundary model and the model that maximized reward fraction per time. We found that predictions from the two models closely matched each other (Figure 4C). Importantly, however, the two models make decisions based on different rules. The constant boundary model makes decisions when the evidence reaches a fixed boundary, which corresponds to a fixed confidence. In contrast, the model that maximizes reward fraction per time does not explicitly assume a decision boundary but decides the number of simulations based on maximizing the simulation efficiency.

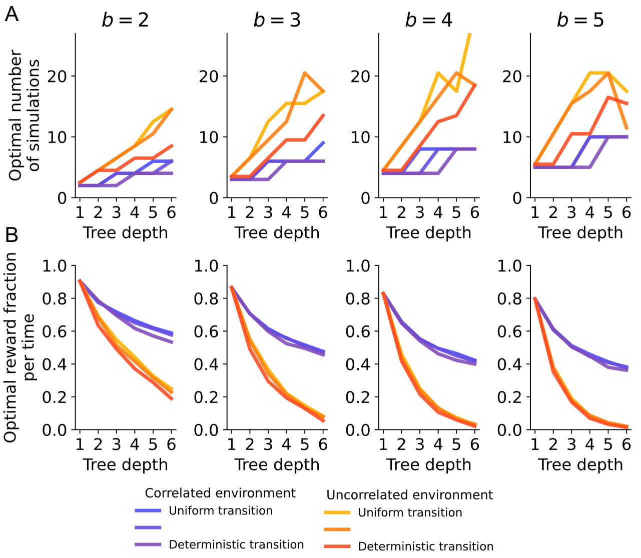

Figure 5: Model performance under various conditions. (A) Optimal number of simulations as a function of tree depth under different branching factors, transition probabilities, and correlation structures. (B) Corresponding optimal reward fraction per time achieved by the agent under different conditions.

When are more rollouts needed?

In the analyses above, we considered a simplified setting with specific assumptions about the environment. It is not clear whether the conclusion that an optimal decision can be based on very few rollouts holds in more general settings, and under what conditions it might fail. To answer this question, we loosened the assumptions and tested our model across a wide range of tree depths, branching factors, and transition distributions. Critically, in the previous analyses, we only considered environments with a correlated reward structure. In these environments with rewards at intermediate nodes, paths leading to leaf nodes closer in the decision tree tend to have more similar cumulative rewards on average. In real-world situations, however, we might also face decision problems without such a correlation structure. In this case, the similarity of cumulative rewards is independent of how close the corresponding nodes are. To investigate the influence of reward correlation structures, we constructed uncorrelated environments where all leaf node rewards were sampled independently and tested our model in both the correlated and uncorrelated environments.

The conclusion that temporally extended decisions can be achieved with few rollouts can fail either when too many samples are necessary to gain a sufficient fraction of rewards, or when rollouts contain no information about expected rewards irrespective of how many samples the agent draws. We therefore looked into the optimal number of simulations and the corresponding optimal reward fraction per time in different settings. We found that in correlated environments, our conclusions held across different branching factors, tree depths, and transition probabilities, demonstrating that the few-rollout strategy is optimal in a wide range of situations (Figure 5A). In uncorrelated environments, however, this picture changes drastically. We found that the agent should perform significantly more rollouts, and the optimal reward fraction per time dramatically dropped when tree depth or branching factor increased (Figure 5B). This was because rollouts contained less information about the expected reward in uncorrelated environments. Thus, the conclusion that an optimal decision can be based on very few rollouts only holds in correlated environments, and the reward correlation structure is a critical factor determining the rollout efficiency.

Discussion

In this paper, we propose a model that can achieve optimal temporally extended decisions with very few rollouts. Our model extends an influential theory by Vul et al. (2014) to the temporally extended setting, suggesting that humans can make optimal decisions based on very few rollouts across a range of environments. Our model provides a normative explanation for when and why humans do not plan to termination. Further, we show that the model can capture previous results that are traditionally attributed to pruning or satisficing strategies without explicitly building in these heuristics, offering an alternative explanation for these results. Finally, we showed that our model has dynamics that closely resemble evidence accumulation models and provides predictions about reaction time patterns in temporally extended decision trees.

Although the number of paths in a decision tree grows exponentially with tree depth, our model predicts that the optimal number of simulations only grows roughly linearly with tree depth in correlated environments (Figure 5). Thus, the time complexity of forward simulation would not cause a demand for an impractical level of resources. We demonstrate that factors including tree depth, branching factor, and transition distribution do not critically affect the efficiency of forward simulation. Rather, the crucial factor that determines the efficiency of forward simulation is the reward correlation structure in the environment.

We assume a constant transition distribution at all non-root nodes, and only the root-node policy is controlled by the agent. Although the actions taken by the agent have temporally extended consequences, this problem is different from the full sequential decision problem. However, this does not mean that our setting is an oversimplification of real-life decision problems. In practice, we often face situations where we make decisions to initiate a series of events that are beyond our control, or where the results of our actions are so uncertain that we cannot reasonably optimize all future decisions. An example is investing in the stock market. We can choose to invest in a particular stock, but we cannot control all future consequences due to random factors such as market fluctuations, economic conditions, company performance, and even global politics. When we make a choice, we are in fact treating the problem as an optimization problem over a single action (e.g., which stock to invest in) under the assumption of an environment with uncertainty. Importantly, realistic environments also often have correlations, e.g. a recession can lead to many possible related futures where we lose money on our investments.

Nevertheless, our assumption of constant transition distributions could introduce disparities when compared to full sequential decision problems. For example, if the agent performs deep rollouts, they should lead to changes in the decision policy at nodes later in the tree. This will change the expected value of actions at the root node and make early rollouts no longer useful. Thus, we speculate that changing the assumption of constant transition distributions would lead to a requirement for more rollouts. Nevertheless, this does not necessarily result in an impractical optimal number of rollouts in correlated environments, since the time cost can eventually overtake the benefit as the agent conducts more rollouts. Future work could generalize our analyses to full sequential decision settings, where the agent has control over its policy at every node.

In addition, we only considered rollout-based forward simulations. While rollouts might be an important component of planning (Jensen, Hennequin, & Mattar, 2023), the exact search algorithms adopted by humans are likely task-specific and more structured than simple rollouts (Callaway et al., 2022). Our model is, therefore, not intended as a complete description of human planning. Nonetheless, little is known regarding the efficiency of these search algorithms when considering the time costs associated with the search process. It is also unclear whether and how humans could adaptively adjust their search algorithms according to the time costs to achieve high efficiency. Our work thus provides insights and establishes a framework for evaluating potential search algorithms adopted by humans. Future work might examine the efficiency of more structured search algorithms to shed light on how mental simulation can effectively support decision-making.

References

Bakkour, A., Zylberberg, A., Shadlen, M. N., & Shohamy, D.

(2018). Value-based decisions involve sequential sampling from memory. *BioRxiv*, 269290.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, *6*(8), 1112–1125.

Erev, I., & Roth, A. E. (2014). Maximization, learning, and economic behavior. *Proceedings of the National Academy of Sciences*, *111*(supplement_3), 10818–10825.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.*, *30*, 535–574.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological science*, *15*(8), 534–539.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, *13*(12), 517–523.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, *8*(3), e1002410.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., … Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098–3103.

Jensen, K. T., Hennequin, G., & Mattar, M. G. (2023). A recurrent network model of planning explains hippocampal replay and human behavior. *bioRxiv*, 2023–01.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

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

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological review*, *119*(1), 120.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Zylberberg, A., Bakkour, A., Shohamy, D., & Shadlen, M. N. (2024). Value construction through sequential sampling explains serial dependencies in decision making. *bioRxiv*, 2024–01.