Resource-Rational Encoding of Reward Information in Planning

Zhuojun Ying (z5ying@ucsd.edu)
Department of Cognitive Science, University of California, San Diego

Frederick Callaway (fredcallaway@nyu.edu)
Department of Psychology, New York University

Anastasia Kiyonaga (akiyonaga@ucsd.edu)
Department of Cognitive Science, University of California, San Diego

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

Abstract

Working memory is widely assumed to underlie multi-step planning, where representations of possible future actions and rewards are iteratively updated before determining a choice. But most working memory research focuses on a context where stimuli are presented simultaneously and the value of encoding each stimulus is independent of others. It is unclear how working memory functions in planning scenarios where the rewards of future actions unfold over time, are retained in working memory, and must be integrated for plan selection. To bridge this gap, we adapted a version of the "mouselab task" in which participants sequentially observe the reward at each node in a decision tree before selecting a plan that maximizes cumulative rewards. We specified a theoretical model to characterize the optimal encoding and maintenance strategy for this task, given working memory constraints, which trades off the cost of storing information with the potential benefit of informing later choices. The model encoded rewards in choice-relevant plans more often, in particular, rewards on the best and (to a lesser extent) worst plans. We then tested human participants, who showed the same pattern in the accuracy of their explicit recall. Our study thus establishes an empirical and theoretical foundation for models of how people encode and maintain information during planning.

Keywords: working memory; planning; information theory; reinforcement learning

Introduction

Imagine that you're at a music festival, trying to plan your day. There are two stages, and it takes a while to walk between them, so you want to avoid switching too much, while also seeing the best music. (This example will work best if you try to follow along!) For the first set, you can see Post Animal at Stage 1 or Polo & Pan at Stage 2. You definitely want to see Polo & Pan, so you'll start at Stage 2. If you stay there for the second set, you can see Bonobo; they're pretty good, sure. But the third set is The 1975? Ugh—no thanks...

This example illustrates how quickly working memory (WM) resources can become overloaded while planning, and also how people can strategically allocate those resources. Without looking back, do you remember all the bands who played at Stage 2? If the example worked, you would remember that Polo & Pan played first and that The 1975 played third, but maybe not that Bonobo played second. Polo & Pan you remember because you chose to start at Stage 2 to see them. On the other hand, you remember The 1975 because you specifically tried to *avoid* seeing them.

A critical role for WM in planning has long been acknowledged (Pribram et al., 1960). For example, Miller et al. (2017) and Owen (2004) posit that both the construction and evaluation of plans occur in WM, and neural data show that manual action plans are integrated into WM concurrently with the initial encoding of visual stimuli (Boettcher et al., 2021). Furthermore, planning is often modeled as the construction of a decision tree (Huys et al., 2012; van Opheusden et al., 2023), a structure which intuitively would have to be stored in WM, since it has to be continually and rapidly updated. Surprisingly however, although the *time* costs of constructing decision trees have motivated many heuristic and normative models of approximate planning (Callaway et al., 2022; Keramati et al., 2011; Sezener et al., 2019), WM constraints are only rarely integrated into computational models of planning (although see MacGregor et al., 2001), and to our knowledge there are no normative models of how an optimal planner would navigate these constraints.

On the other hand, outside of a planning context, there is an immense body of theoretical work formally characterizing WM capacity and optimal strategies for deploying it (e.g., Barlow, 1961; Stocker and Simoncelli, 2006; Sims, 2016). Many of these models rely on the mathematics of *information theory*. The key idea is to view WM as a capacity-limited information channel or "bottleneck", which restricts the amount of information that can pass from sensory input into internal representations, and ultimately actions. Although much of this work focuses on our ability to veridically reconstruct stimuli (van den Berg and Ma, 2018; Sims et al., 2012), these formal tools have also been applied to understand reward-driven choice, where the goal is not to simply remember what you have seen, but to instead apply the information to select reward-maximizing actions (Sims, 1998; Bhui and Gershman, 2018; Matêjka and McKay, 2015).

However, while these models have occasionally been applied in sequential decision making contexts (e.g., Ortega and Braun, 2013), they have not been applied to the problem of planning itself, specifically the sequential evaluation of different possible courses of action. Most planning models assume that the outcomes of action evaluations are encoded perfectly during planning and are not influenced by the order of action evaluation or the precision of other encoded action values.

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.*

This assumption likely overestimates the information available to humans during planning.

Here, we seek to develop an optimal information-theoretic model of planning under WM constraints. This goal faces two key challenges. First, information-theoretic models are almost always one-shot, in the sense that a single set of stimuli is presented all at once, jointly encoded, and then decoded into an action (see Woodford, 2016 for a notable exception). In contrast, planning involves iteratively considering different pieces of information and integrating them into a dynamic representation (for example, adding nodes to a decision tree). Second, these models often assume that the importance of each stimulus is independent of the other stimuli (or dependent in a simple way, for example if you only need to remember the most rewarding action in a set). In contrast, planning involves recursive maximization over multiple actions; this introduces complex dependencies, such that the relevance of one action depends critically on the actions that could be taken before or after it.

To address these challenges, we draw on two recently developed formal tools. First, we model planning within the metalevel Markov decision process framework (metalevel MDPs; Callaway et al., 2022). This allows us to develop an optimal model that respects the sequential nature of planning. Second, we draw on work in artificial intelligence that formalizes information-theoretic bounds (and optimal solutions) for problems where an agent receives information sequentially (Fox and Tishby, 2012). Combining these two formalisms, we can formalize the problem of planning under WM constraints, and characterize its optimal solution.

Planning with limited working memory

Here, we formalize the problem of decision-tree search under information-theoretic WM bounds. In particular, we aim to characterize the optimal strategy for encoding and maintaining information about rewards in a decision tree. For simplicity and tractability, we focus exclusively on this aspect of the problem. That is, we do not consider the problem of which future states to consider, nor the problem of when to stop planning and make a choice (the optimal solution to these problems has been characterized by prior work; Callaway et al., 2022). We thus assume an exhaustive search strategy in which future states are considered one by one in a depth-first manner.[1] Extending the model to account for how WM constraints interact with the search strategy is an important (and challenging) direction for future work.

We begin with an intuitive summary of the model, providing details of the formalization and solution strategy below. On each trial or *episode*, the agent is presented with a decision tree of known size but unknown rewards (intuitively, a stable spatial environment but variable goals). The goal is to select a sequence of actions that maximizes total reward. Before taking action, the agent systematically considers every state, taking note of the reward there and storing it in a decision tree. Absent WM constraints, the agent could then trivially identify the optimal plan and perfectly maximize reward. However, the agent has limited WM, and thus pays a cost associated with storing information about the rewards. Therefore, after considering each reward, the agent makes a choice about how precisely they will maintain the rewards into the next time step. This choice applies to the full set of rewards considered thus far; that is, the agent could initially pay the cost to maintain a reward with high precision, but decide to forget about it several timesteps later. This decision is permanent, in the sense that once information about a reward is lost, it cannot be recovered. Once all the rewards have been considered, the agent selects an action sequence based on their final mental state.

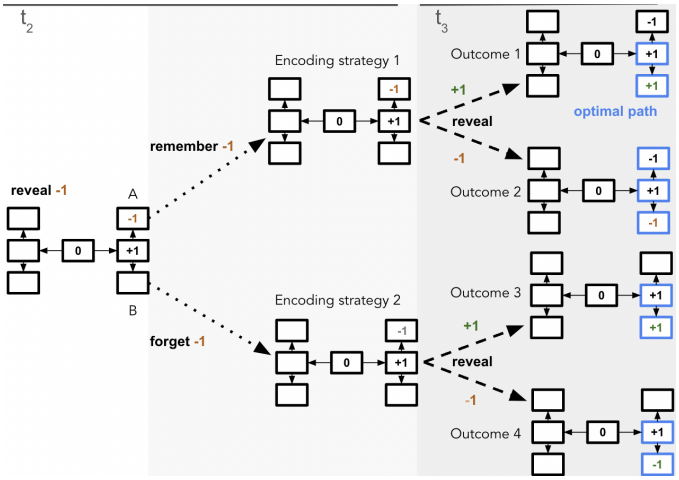

Figure 1: Binary decision tree reward representation. The white panel shows the current information available at timestep 2. The light grey panel shows two possible encoding strategies at timestep 2. The dark grey panel shows the reward representation outcome at timestep 3, based on the encoding strategies from timestep 2. Dotted lines indicate encoding strategies at timestep 2, while dashed lines indicate possible rewards revealed at timestep 3. Reward values are color-coded: +1 in green, -1 in orange. Optimal plans are marked in blue.

Despite its apparent simplicity, this task poses a non-trivial sequential decision problem. For example, consider the example shown in Figure 1, a simple binary decision tree with binary rewards, in which the agent has discovered a mediocre plan with rewards +1 and -1. If the agent were to only receive the information available at timestep 2 in the white panel, plan A would be valued at zero, and plan B at +1. In this scenario, the agent should remember the +1 to ensure selection of a plan in the right branch, and also remember the -1 to choose plan B over plan A (i.e., encoding strategy 1 in the light grey panel). However, this approach may not be optimal when considering the reward that will be revealed at $t = 3$ in the dark grey panel. If a reward of +1 is revealed (i.e.,

[1] That is, all future states along one plan are considered before future states in another plan are evaluated.

outcomes 1 and 3), plan B becomes the optimal choice, suggesting that only the nodes on plan B need to be encoded. Conversely, if the reward at $t = 3$ is -1 (i.e., outcomes 2 and 4), plans 1 and 2 would have equivalent values, making it unnecessary to remember that both nodes have a value of -1. Therefore, when future information is taken into account, the optimal strategy would be to encode only the +1 value (i.e., encoding strategy 2), as this decision accounts for all the potential rewards to be revealed at the next timestep. This example thus illustrates how accounting for future information can change what is worth storing in WM.

Problem formulation

Following Callaway et al. (2022), we specify the model as a metalevel Markov decision process. We thus formalize the problem of WM-bounded planning in terms of an agent selecting mental operations to update their mental state in order to balance cognitive cost with external reward. However, in contrast to Callaway et al., which focuses on the decision of *which* rewards to consider, we instead consider the decision of *how precisely* to maintain information about the rewards that have already been considered. Similarly, we put aside the cost associated with the initial evaluation of reward and instead focus on the cost associated with maintaining information about rewards over multiple time steps.

The model is defined by a set of possible world states, a set of mental states (noisy representations of the world state), and a set of mental actions that update the mental state. We further define cost and value functions over mental states that capture the cost of maintaining a representation in WM and the expected reward associated with selecting actions using a given mental state. We describe each component below.

World state The true state of the world specifies the actual reward at each future state. It is represented by a vector W, where $W^{(k)}$ denotes the reward at state k. The world state is constant across all time steps and is hidden from the agent. However, the reward at each state is drawn from some distribution $p(W^{(k)})$ that is known to the agent, providing a prior belief absent any specific information about a given reward.

Mental state At each time step t, the agent holds a mental state M_t, a noisy representation of the rewards, W. For simplicity, we assume that the represented rewards are corrupted by Gaussian noise, so that $M_t^{(k)} \sim \mathcal{N}(w^{(k)}, \sigma_t^{(k)})$[2]. The degree of noise for each reward $\sigma_t^{(k)}$ is controlled by the agent, as described below. Implicitly, the agent also tracks the precision with which they have maintained each reward (the $\sigma_t^{(k)}$).

Mental action At each time step, the agent executes a mental action that controls the precision with which each reward is represented. Concretely, a mental action sets $\sigma_t^{(k)}$ with two restrictions. First, rewards that have not been considered yet have zero precision: $\sigma_t^{(k)} = \infty \; \forall \; k > t$ (we assume that nodes are considered in index order). Second, encoding precision can never increase: $\sigma_t^{(k)} \geq \sigma_{t-1}^{(k)} \; \forall \; k < t$. Thus, the agent can encode the currently considered reward $w^{(t)}$ at any precision, and can lower the precision of previously encoded rewards.

Mental cost function At each time step, the agent pays a cost associated with the reward information maintained in WM. We quantify this cost using mutual information, the amount of information obtained about the world state W when observing the mental state M_t. The cost of M_t is higher if M_t more veridically represents the actual rewards and contains a larger amount of information about W. For a given reward, the mutual information is defined

$$\mathrm{I}\left(M_t^{(k)}; W^{(k)}\right) = \int_{w^{(k)}} \int_{m_t^{(k)}} p\left(m_t^{(k)}, w^{(k)}\right) \log\left(\frac{p(m_t^{(k)}|w^{(k)})}{p(m_t^{(k)})}\right). \tag{1}$$

Because we assume independent and identically distributed rewards in both M_t and W, the total mutual information and thus the total cost is simply the sum across all the rewards.

Physical action After all rewards have been considered and encoded, the agent constructs a plan. Because we limit our attention to deterministic problems, the plan can be represented as a sequence of nodes to visit, which we denote τ. The value of a plan given a world state is simply $V(\tau, w) = \sum_{k \in \tau} w^{(k)}$. We assume that the agent chooses a plan that maximizes the expected value given their mental state, that is

$$p(\tau|m) \sim \text{Uniform}\left(\underset{\tau}{\text{argmax}} \sum_{w} p(w|m) V(\tau, w)\right), \tag{2}$$

where $p(w|m)$ is computed by Bayes rule. Intuitively, this means that an imprecisely encoded reward will revert towards the prior mean, assuming that the agent tracks the precision with which they have maintained each reward.

Terminal reward function The terminal reward function captures the expected reward attained from the executed external actions (when planning terminates), marginalizing over both τ and w given the current mental state.

$$R_T(m) = \sum_{w} p(w|m) \sum_{\tau} p(\tau|m) V(\tau, w) \tag{3}$$

Optimal policy

Having specified the model, we now discuss our strategy for approximating its optimal solution. We draw heavily on the approach of Fox and Tishby (2012), using dynamic programming to compute the optimal value function over mental states, and the Blahut-Arimoto algorithm to determine the transitional probability over mental states, which is used in dynammic programming.

Note that we make one key approximation. In the complete model, noise in the representation influences the selection of both physical and mental actions. However, for tractability, we apply a mean field approximation with respect to the latter.

[2]Here $w^{(k)}$ and $m^{(k)}$ represent individual instances of the actual and noisy representation of reward at state k, respectively

This allows us to convert the problem of selecting encoding precisions $\sigma_t^{(k)}$ given a specific noisy representation m_t into the much easier problem of directly identifying the optimal feasible mental state distribution given the current one and the observation: $p(M_t|M_{t-1}, w^{(t)})$. Due to space constraints, we refer the reader to Fox and Tishby (2012) for details. We give a brief summary below.

The solution is identified with dynamic programming, using a Bellman equation defined over mental states,

$$V(M_t) = R(M_t) + \gamma \sum_{M_{t+1}|M_t} p(M_{t+1}|M_t) V(M_{t+1}), \quad (4)$$

where γ is a discount factor and the reward function is defined

$$R(M_t) = \mathbf{1}(t = T) R_T(M_t) - \lambda \sum_{w|M_t} p(w) \mathrm{I}(M_t; w), \quad (5)$$

with λ defining the tradeoff between value and information.[3] That is, the agent pays an information cost at every time step and receives the termination reward at the final time step.

The policy soft-maximizes value $V(M_t)$, selecting from the set of possible mental states. Given M_{t-1} and $w^{(t)}$, the possible mental states are the ones where the mean of $M_t^{(k)}$ is $w^{(k)}$ (for all $k \leq t$, zero otherwise) and the standard deviations follow the constraints defined above under "Mental action". Note that the means are fixed to the true reward because of the mean-field approximation discussed above.

Finally, we can define a transition function directly over mental states by marginalizing over $w^{(t)}$:

$$p(M_t|M_{t-1}) = \sum_{w^{(t)}} p(M_t|M_{t-1}, w^{(t)}) p(w^{(t)}). \quad (6)$$

Note that the policy depends on the value function, which in turn depends on the transition function $p(M_t|M_{t-1})$, which itself depends on the policy. Fortunately, this cyclic dependence can be resolved by the Blahut-Arimoto algorithm, which alternates between updating the policy and the marginal until convergence.

Experiment

To empirically test how well our model characterizes human encoding strategies, we designed an experiment that mirrors the complexities of real-world planning. In this experiment, participants were presented with a sequence of choices, each with associated rewards, simulating the sequential decision-making process.

For tractability, we used a minimalistic version of the task: a two-step binary decision tree and a binary reward distribution (-4 and $+4$ with equal probability.) The experiment was structured as follows: The decision tree was depicted as a game board, the reward at each node was represented by a radial frequency pattern, and plans were represented by paths starting at the central node and ending at a leaf node. Participants were required to sequentially observe the reward at each node and then navigate a plane from the central node to a leaf node, aiming to collect the highest possible cumulative rewards across the visited nodes. Participants were then asked to recall the reward at each node.

Methods

Participants We recruited 70 participants with normal or corrected-to-normal vision through UCSD's Sona system.

Stimuli The primary stimulus in this experiment was a game board designed as a 7-node decision tree (Figure 2.b). The reward at the central node of this board was always set to 0. The rewards at other nodes were visually represented as radial frequency patterns, and each had a value of either -4 or $+4$ with equal probability (Figure 2.a). Participants were trained to associate these patterns with their corresponding reward values before starting the main experiment.

Procedure Each participant completed one practice trial followed by 39 actual trials. To ensure consistency, the sequence of trials and the order of nodes probed during reward recall in each trial were standardized across all participants. After each trial, we revealed the actual decision tree and the assigned rewards to provide feedback on their performance.

Task Each trial consisted of three phases: reward presentation, path selection, and reward recall (shown in Figure 2.c). During reward presentation, participants observed the reward at each node on the board sequentially for 1.5 seconds each without inter-stimulus intervals, following a depth-first manner. Subsequently, during path selection, they were asked to control a plane to travel from the central node to one of the leaf nodes, aiming to accumulate the maximum possible rewards in the nodes they have visited. Once the plane moved, it could not return to a visited node. We incentivized participants' performance in path selection by informing them that the experiment concludes when 200 points of reward have been accumulated across all trials, which corresponded to the maximum number of points possible across all 40 trials. After completing the route, we asked participants to recall the reward at each node. Specifically, they were prompted to match the pattern on a probed node with the corresponding pattern from their memory by adjusting a slider to alter the pattern until it matched their recollection. We probed the participants's recall of the reward at each node in random order. The participants were required to complete the path selection and reward recall phases in the practice trial without errors to proceed to the actual experiment.

Model

Although the theoretical model allows for arbitrary degrees of precision, for tractability, we only considered two possible levels. The agent could either choose to perfectly encode the

[3] A smaller λ indicates that the mental cost is downweighted when evaluating the mental states.

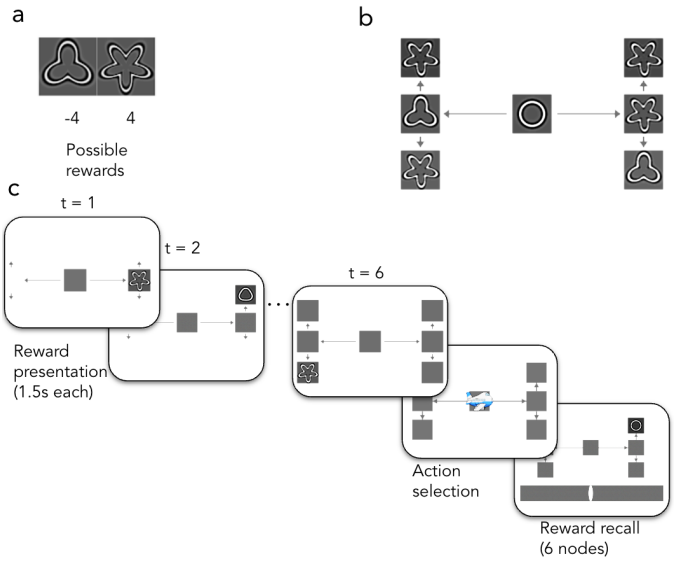

Figure 2: **a:** Radial frequency patterns corresponding to reward values. The possible reward values $\mathcal{R} = \{-4, +4\}$, each with probability 0.5. The radial frequency pattern representing the reward value -4 had 3 "bumps", and the radial frequency pattern representing the reward value $+4$ had 5 "bumps". **b:** Example of a decision tree presented to participants, with the central node fixed at a reward of 0. Other nodes displayed rewards from $\mathcal{R} = \{-4, +4\}$. **c:** Overview of the experimental trial process, with each participant completing 40 trials featuring randomly assigned rewards.

reward information or completely forget it. Furthermore, because the reward distribution was binary, the agent adopted a discrete prior, assigning equal weight to the two possible values (-4 and $+4$). Therefore, if the actual reward is encoded, $M_t^{(k)}$ is a delta distribution on the true reward. Otherwise it reverts to the prior, -4 or $+4$ with equal probability.

Results

Both simulated and behavioral data revealed that better performance in the path selection task is associated with a higher number of nodes encoded in WM. Additionally, our model that incorporated WM constraints was more effective in predicting the participants' path selection than a variant of the model without this constraint. Furthermore, we found that both the model and human participants employed similar encoding strategies, encoding reward information at leaf nodes in the most and least favorable paths more often than leaf nodes in paths with the value 0. Detailed analyses of these findings are presented in the sections below.

Tradeoff Between Value Maximization and Information Minimization Our model explored the balance between maximizing the value of the selected path and minimizing WM load, regulated by the λ parameter. An increase in λ leads to a lower average reward of selected path, as shown in Figure 3.a. In our simulations, there was a significant negative correlation between the average value of the chosen path and the average number of forgotten nodes across different λ values (Pearson's correlation coefficient $r = -0.86$, $p < .001$). This significant negative correlation was also observed in our behavioral data ($r = -0.68$, $p < .001$, shown in Figure 3.b).

This result indicated that human participants showed a trade-off between the performance in planning and the amount of information that was stored in WM. Note that although the simulation results spanned a wide range of possible λ values, the number of forgotten nodes for humans consistently remained at the lower end of this range. This observation may suggest that humans prioritize performance over reducing WM load during this task.

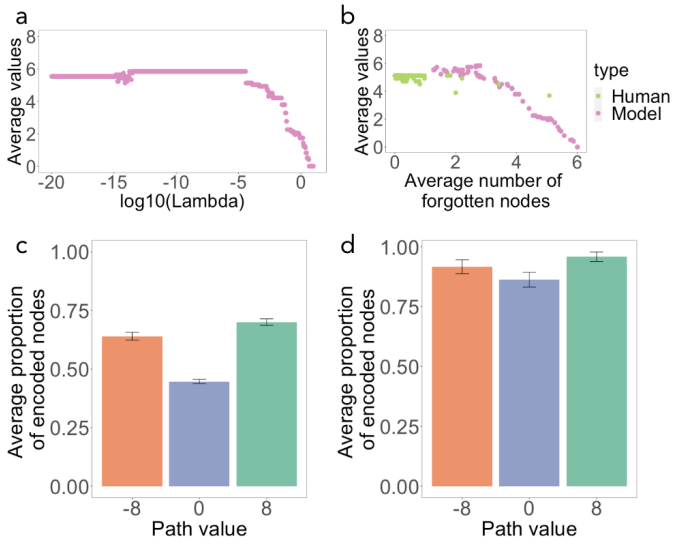

Figure 3: **a:** Relation between 700 λ values (ranging from $1e-20$ to 10) with constant $\gamma = 0.5$ and the average reward of selected path per agent across the 39 actual trials presented to the participants (smaller λ indicates the mental cost is downweighted when evaluating mental actions). **b:** Average number of forgotten nodes per trial vs. average value of the selected path for the model (purple) and human participants (green), one dot per agent/participant. Here, the path selected by the model was the one with maximal expected reward given the mental state (ties were broken randomly). **c:** Simulation showing the proportion of encoded leaf nodes in paths with reward values -8, 0 and 8, across 700 λ values (from $1e-20$ to 10) at a constant $\gamma = 0.5$. **d:** Experimental data showing the proportion of encoded leaf nodes in paths with reward values -8, 0 and 8.

Path Selection Prediction We evaluated the performance of our model by fitting the parameters λ and γ based on the participants' reward recall responses and using the fitted parameters to predict the participant's path selection. We used the BOBYQA algorithm (Powell et al., 2009) with maximum likelihood estimation for paramter fitting. We used the first

20 trials for parameter fitting and the last 19 trials for path selection prediction. Spearman's correlation confirmed a significant positive correlation between the participants' reward recalls and the predicted reward representations (Spearman's $\rho = 0.84$, $p < .001$). Predictions of selected paths based on these representations yielded an accuracy of 58.2%, which exceeded the 25% accuracy expected by chance. These results showed that we were able to predict the participants' selection of paths from the amount of WM resources allocated to reward information encoding during planning, which was estimated by the parameter λ.

We then used cross validation with 2 folds to compare between 3 models:

- our nornative model under WM constraints,
- a variation of our model without WM constraints ($\lambda = 0$ and only fitting γ),
- and a model that randomly chose to encode the reward information at each node or not.

Our model with WM constraints had a log likelihood of -1546.25, the model without WM constraints had a log likelihood of -1588.45, and the random model had a log likelihood of -1925.25.

The participants' encoding strategies could be better characterized by the model with WM contraints, compared with a model without, or a model that randomly chose which nodes to encode. This result indicated that human participants did not encode the rewards at all nodes perfectly, and this limitation influenced their path selection. However, the performance of the model with and without WM constraints was similar. One explanation might be that this task did not heavily tax the participants' WM, as evidenced by the small number of forgotten nodes for humans shown in Figure 3.b.

Reward Encoding Strategy We posited that the likelihood of a node being encoded depends on its relevance in future path selection. To test this hypothesis, we ran simulations with 700 λ values ranging from $1e-20$ to 10 and constant $\gamma = 0.5$. Consistent with this hypothesis, our simulations at the final timestep revealed that leaf nodes in paths with the most extreme values (-8 or 8) were encoded more frequently than other leaf nodes. Specifically, Mann-Whitney U tests showed that paths with the value 0 had a smaller proportion of encoded leaf nodes (0.45, SD = 0.15) compared to paths with the value -8 (0.64, SD = 0.27; $p < .001$) and paths with the value 8 (0.70, SD = 0.22; $p < .001$). Paths with the value -8 had a smaller proportion of encoded leaf nodes compared to paths with the value 8 ($p < .001$), as shown in Figure 3.c.

We observed a similar pattern in our behavioral experiment[4]. Paths with the value 0 had a smaller proportion of encoded leaf nodes (0.86, SD = 0.15) compared to paths with the value -8 (0.92, SD = 0.15; $p < .001$) and paths with the value 8 (0.96, SD = 0.10; $p < .001$). Additionally, paths with the value -8 had a smaller proportion of leaf nodes compared to paths with the value 8 ($p = 0.002$), as shown in Figure 3.d.

Since the absolute values of the reward at each node were consistent, this result implied that in our task, whether a participant chose to encode a node or not was not dependent on the nodes' individual reward values. Instead, the participants considered the importance of a piece of reward information in the context of finding the optimal path multiple timesteps after they observed this information. This indicated that during planning, the potential of a node to aid in future decisions was considered more critical than its immediate value.

Discussion

In this study, we presented a information-theoretic model that encoded reward information strategically during planning to maximize the value of future plans under WM constraints. We hypothesized that reward information related to plan selection will be encoded with higher probability. Our simulation and behavioral data confirmed this hypothesis.

Both our model and the human participants encoded leaf nodes in paths with extreme values more often compared to leaf nodes in paths with the value 0, despite the constant absolute value of node rewards. This implied that during reward information encoding, whether a node can help in path selection was considered more important than its immediate value. This finding aligned with prior research showing that WM prioritizes goal-related items and encodes them with greater precision (Hu et al., 2016; Ravizza et al., 2021). Our results extended this finding to the context of two-step planning, indicating that action-relevant information in goal-related plans is encoded with higher precision in WM. However, this result could also be explained by varying cognitive demands: paths valued at 0 contained stimuli with different shapes, requiring participants to associate each shape with its position. Conversely, paths with extreme rewards contained identical stimuli, which may reduce cognitive load.

Our analysis was confined to a simple binary decision tree with binary rewards, alongside a binary encoding decision – either encoding a reward or not. This approach may limit the generalizability of our model, as existing WM research indicates that items in WM can be encoded with varying degrees of precision (Sims et al., 2012). Future research could explore encoding strategies for actions with continuous rewards and might consider a model where precision levels are continuous rather than binary.

In conclusion, this study showed that the value of future actions were not stored in WM with equal probability, and the probability of encoding an action value is related to the relevance of the action in plan selection. Future work should investigate how humans decide which future actions to evaluate during planning, given the constraints of WM, and examine how the combination of both action evaluation and encoding strategies influence the planning outcome.

[4]A node was considered remembered if its reported reward value shared the same sign as the actual reward because there were only 2 possible reward values.

References

Barlow, H. B. (1961). Possible Principles Underlying the Transformations of Sensory Messages. In Rosenblith, W. A., editor, *Sensory Communication*, page 0. The MIT Press.

Bhui, R. and Gershman, S. J. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*, 125(6):985–1001.

Boettcher, S. E., Gresch, D., Nobre, A. C., and van Ede, F. (2021). Output planning at the input stage in visual working memory. *Science Advances*, 7(13):eabe8212.

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

Fox, R. and Tishby, N. (2012). Bounded planning in passive pomdps. *arXiv preprint arXiv:1206.6405*.

Hu, Y., Allen, R. J., Baddeley, A. D., and Hitch, G. J. (2016). Executive control of stimulus-driven and goal-directed attention in visual working memory. *Attention, Perception, & Psychophysics*, 78:2164–2175.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and 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.

Keramati, M., Dezfouli, A., and Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLOS Computational Biology*, 7(5):e1002055.

MacGregor, J. N., Ormerod, T. C., and Chronicle, E. P. (2001). Information processing and insight: a process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):176.

Matêjka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298.

Miller, G. A., Eugene, G., and Pribram, K. H. (2017). Plans and the structure of behaviour. In *Systems Research for Behavioral Science*, pages 369–382. Routledge.

Ortega, P. A. and Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683.

Owen, A. M. (2004). Cognitive planning in humans: New insights from the tower of london (tol) task. In *The cognitive psychology of planning*, pages 145–162. Psychology Press.

Powell, M. J. et al. (2009). The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26.

Pribram, K. H., Miller, G. A., and Galanter, E. (1960). Plans and the structure of behavior. *New York*.

Ravizza, S. M., Pleskac, T. J., and Liu, T. (2021). Working memory prioritization: Goal-driven attention, physical salience, and implicit learning. *Journal of Memory and Language*, 121:104287.

Sezener, C. E., Dezfouli, A., and Keramati, M. (2019). Optimizing the depth and the direction of prospective planning using information values. *PLOS Computational Biology*, 15(3):1–21.

Sims, C. A. (1998). Stickiness. *Carnegie-Rochester Conference Series on Public Policy*, 49:317–356.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, 152:181–198.

Sims, C. R., Jacobs, R. A., and Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological review*, 119(4):807.

Stocker, A. A. and Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585.

van den Berg, R. and Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *eLife*, 7:e34963.

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., and Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, pages 1–6.

Woodford, M. (2016). Optimal Evidence Accumulation and Stochastic Choice. *Working Paper*, pages 1–46.