Extended Metalevel Planning with Reinforcement Learning under Cognitive Constraints

LiYingjun (ID: 21307099)

14. januar 2025

Resumé

Human planning often occurs under constrained cognitive resources. Recent work by Callaway et al. (2023) [1] has formalized planning as a metalevel Markov Decision Process (MDP), modeling the internal operations of planning and their associated costs. In this paper, we extend their framework, focusing on the *constant-variance* condition in Experiment 2 and incorporating a cognitively controlled component to explore how individuals adapt their planning strategies. Collaborating with a senior student from the Psychology department, we investigate whether human planners align with rational metareasoning predictions when resource limitations, cost structures, and environment uncertainties are carefully manipulated.

We further expand on how Reinforcement Learning (RL) principles, especially modern deep RL algorithms, can be integrated with metalevel MDP formulations. Our preliminary codebase, accessible at `https://github.com/KFCCrazzzyThursday/optimal-planning-algorithms`, implements a scalable RL approach that approximates optimal planning strategies under varying computational costs. Some of the data used in this experiment was provided by the original author, Fred Callaway, under an agreement that prevents public release at this time.

Nonetheless, our preliminary analyses indicate that human participants balance exploration and exploitation in a manner akin to RL agents, lending support to the idea that metalevel planning is effectively a sequential decision-making process at the cognitive level.

Introduction

Planning is a core aspect of human cognition, yet it is inevitably subject to resource constraints in terms of time, working memory, and mental effort. While traditional planning algorithms, such as A* or dynamic programming, often assume unbounded computation and perfect foresight, empirical studies highlight that people do not exhaustively explore all future possibilities. Instead, they adaptively prune search branches, revise decisions mid-process, and converge on *reasonably* good choices with bounded rationality.

A growing body of research frames planning as a metalevel Markov Decision Process (MDP) [1], wherein the process of planning itself is treated as a sequential decision problem. The agent must decide which cognitive operations (e.g., node expansions in a decision tree) to perform and when to stop planning, balancing cognitive cost against the expected improvement in outcome quality. In simpler terms, we can view it as a resource-constrained cost-benefit analysis in which each computation carries a penalty, but may yield better final decisions.

This paper builds on the work of Callaway et al. (2023) who introduced Mouselab-MDP, a paradigm that externalizes human planning steps via clickable states. We specifically focus on the *constant* condition of Experiment 2, which the authors designed to test whether participants adjust their search depth or branching according to stable reward distributions. By collaborating with a senior researcher in Psychology, we further explore how *cognitive control* mechanisms might modulate these planning strategies, extending the original model to account for variations in participants' metacognitive awareness and self-regulatory processes.

Background and Related Work

Rational Metareasoning and Metalevel MDPs

Rational metareasoning posits that an intelligent agent must continuously solve a "metalevel" problem: how to efficiently allocate computational resources to subproblems in pursuit of better decisions. This perspective is especially relevant in multi-step decision tasks, where enumerating the entire search tree can be infeasible. Formalizing planning within a metalevel MDP [2] typically involves:

- **Metalevel States (Beliefs)**: Representing partial knowledge or expansions of the decision tree.
- **Metalevel Actions (Computations)**: Operations such as node expansion, which reveal reward or cost information about a potential future state.
- **Rewards and Transition Dynamics**: Balancing the internal cost (time, effort) of additional searches against the external reward gained from executing a better-informed plan.

While classical MDPs focus on action selection in an external environment, metalevel MDPs direct attention inward, modeling the "state of knowledge" and "cognitive effort" as part of the decision-making loop.

Human Planning under Cognitive Constraints

Empirical findings in cognitive psychology show that people do not purely maximize expected reward in complex tasks, but instead adapt their search depth, direction, and termination conditions [3]. The Mouselab-MDP paradigm, introduced in prior work, traces how participants reveal occluded rewards across multiple steps [1]. In Experiment 2, the authors manipulated the variability (variance) in reward distributions across different depths of a decision tree. The "constant" condition implied that each node in the planning tree had similarly distributed rewards, in contrast to "increasing" or "decreasing" variance conditions. Their results indicated that participants occasionally deviated from an ideal best-first or breadth-first search, underscoring the need to factor in individual differences in cognitive constraints.

Reinforcement Learning Approaches

Reinforcement Learning (RL) methods have shown promise in automating search and decision processes under uncertainty. In standard RL [4], an agent interacts with an external environment, learning a policy to maximize cumulative rewards. By contrast, *metalevel* RL shifts the focus to internal computations as the environment, allowing the agent to learn when and how to plan. Recent advances in Deep RL—such as Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), and Soft Actor-Critic (SAC)—provide scalable approaches to high-dimensional problems. Integrating these with metalevel MDPs raises interesting possibilities: an adaptive system could "learn to plan" by approximating the mapping from partial knowledge states to the best next computational action, making the resource allocation strategy itself the subject of reinforcement learning.

Problem Definition and Research Aim

Resource-Constrained Planning as a Metalevel MDP

Following Callaway et al. (2023), we consider a situation where an agent must balance the external rewards of its plan against the cognitive cost of each planning operation. Specifically, each *metalevel state* captures partial information about the decision tree (i.e., which nodes have been revealed), and each *metalevel action* either expands one node of the tree or terminates planning to execute the best available plan. Formally, this metalevel decision-making process can be represented by a tuple

$$(\mathcal{B}, \mathcal{C}, P_m, R_m, \gamma),$$

where $\mathcal{B}$ is the set of possible belief states, $\mathcal{C}$ is the set of cognitive actions, P_m is the transition function governing how expansions update the belief state, R_m is the metalevel reward function, and γ is a discount factor (used if future metalevel returns are to be discounted).

Belief States and Frontier. Each belief state $\mathbf{s} \in \mathcal{B}$ can be viewed as a vector of length N, where N is the total number of nodes in the underlying environment's decision tree. The entry s_i is either the revealed reward associated with node i or a placeholder value (e.g., $\varnothing$) indicating that node i has not been expanded yet. A node is *expandable* if it is unexpanded ($s_i = \varnothing$) *and* its parent is already revealed. Formally, the set of allowable expansion actions at state $\mathbf{s}$ is given by

$$\text{frontier}(\mathbf{s}) \;=\; \Big\{\, a_i \,\Big|\, s_i = \varnothing \;\wedge\; \text{parent}(s_i) \neq \varnothing \Big\}. \tag{1}$$

Choosing a_i expands node i, revealing its reward from a node-specific distribution.

Metalevel Transition Function. When the agent executes an expansion action a_i, the new belief state $\mathbf{s}'$ is identical to $\mathbf{s}$ except that s'_i now contains a sampled reward from the appropriate distribution. A special *termination* action, often denoted by $\perp$, leads to a terminal metalevel state in which the agent stops further expansions and proceeds with a final plan.

Metalevel Reward Function. The metalevel reward $r(\mathbf{s}, a)$ accounts for both the cost of expanding nodes and the eventual external payoff. Formally, let λ be the cost to expand any single node:

$$r(\mathbf{s}, a) \;=\; \begin{cases} \max\limits_{p \in P} V(\mathbf{s}, p), & \text{if } a = \perp, \\ \\ -\lambda, & \text{otherwise.} \end{cases} \tag{2}$$

Here, taking the termination action ($a = \perp$) yields the maximal expected value over all *complete plans* p, where each plan p is a path from the current state to a terminal state. The expected value of executing plan p in belief state $\mathbf{s}$, denoted $V(\mathbf{s}, p)$, sums over the (revealed or expected) rewards of the nodes in p:

$$V(\mathbf{s}, p) \;=\; \sum_{i \in p} \begin{cases} \mathbb{E}[R_i], & \text{if } s_i = \varnothing, \\ s_i, & \text{otherwise.} \end{cases} \tag{3}$$

Thus, expanding a node incurs a fixed negative reward $-\lambda$, while terminating planning grants whatever value is estimated to be best, given the (partially) revealed subtree. This framework captures the essence of resource-constrained planning: each additional computation carries a cost, yet may improve the final decision quality.

Model Specifications: Stopping and Selection Rules

We compare multiple parameterized models, each defining a *metalevel policy* as a distribution over actions, $\pi(a \mid s)$. Following prior work, we assume that on each decision step, the policy is sampled through the four-step process below. Let frontier(s) denote the set of unexpanded nodes in state s whose parents have been expanded (i.e., the allowable node expansions).

1. **Frontier Check**: If no nodes remain unexpanded (frontier$(s) = \varnothing$), the model must terminate planning (i.e., choose the termination action $\perp$).
2. **Random Action**: If frontier(s) is not empty, then with probability ε the model selects one allowable expansion at random.

3. **Stopping Rule**: Otherwise (with probability $1 - \varepsilon$), the model chooses to stop (i.e., choose $\perp$) with probability $p^M_{\text{stop}}(s)$, which is *model-dependent.* If it stops, planning ends and the agent executes the current best plan.

4. **Selection Rule**: If the model does not stop, it must choose one node $a \in \text{frontier}(s)$ to expand. The probability of selecting node a is given by $p^M_{\text{select}}(s,a)$, also dependent on the model's specification.

Heuristic Models. The "best-first," "depth-first," and "breadth-first" models share a common *stopping* mechanism that blends *absolute* and *relative* satisficing criteria. Let V_{best} be the highest expected value found so far and V_{next} the second highest. Then the stopping probability is parameterized by a logistic function of a linear combination:

$$p^H_{\text{stop}}(s) \;=\; \frac{1}{1 \;+\; \exp\Bigl\{-f_{\text{stop}}(s)\Bigr\}}, \tag{4}$$

where

$$f_{\text{stop}}(s) \;=\; \beta_{\text{satisfice}} \cdot V_{\text{best}} \;+\; \beta_{\text{bestnext}} \cdot \Bigl(V_{\text{best}} - V_{\text{next}}\Bigr) \;+\; \theta_{\text{stop}}. \tag{5}$$

Here, $\beta_{\text{satisfice}}$ and β_{bestnext} control how absolute and relative thresholds affect the slope of the logistic function; θ_{stop} shifts its midpoint. Setting one parameter large and the other to zero can yield a "hard" satisficing policy based on a fixed aspiration level.
For *selection*, each heuristic model approximates a classical search strategy:

$$p^H_{\text{select}}(s,a) \;=\; \mathbf{1}\Bigl(a \in \text{frontier}(s)\Bigr)\, \frac{\exp\Bigl\{\beta_{\text{select}} \cdot f^{\text{ALG}}_{\text{select}}(s,a)\Bigr\}}{\sum_{a' \in \text{frontier}(s)} \exp\bigl\{\beta_{\text{select}} \cdot f^{\text{ALG}}_{\text{select}}(s,a')\bigr\}}, \tag{6}$$

where $f^{\text{ALG}}_{\text{select}}$ is a *node-scoring function* specialized to the particular heuristic:

$$f^{\text{BEST}}_{\text{select}}(s,a_i) = V(s,i), \quad f^{\text{DEPTH}}_{\text{select}}(s,a_i) = \text{depth}(s,i), \quad f^{\text{BREADTH}}_{\text{select}}(s,a_i) = -\,\text{depth}(s,i).$$

Hence, in the limit of large β_{select}, the model deterministically emulates the chosen strategy.

Random Model. A "random" policy takes the same form but sets $\varepsilon = 0$ and uses a constant logistic function for stopping, i.e., $f_{\text{stop}}(s) = \theta_{\text{stop}}$. Its selection score is zero for all nodes, so expansions are chosen uniformly at random whenever planning continues.

Optimal and Myopic Models. Finally, the "optimal" model defines stopping and selection in terms of the *optimal state-action value function* Q_λ of the underlying metalevel MDP with computational cost λ. Let β_{stop} and β_{select} be inverse-temperature parameters. Then the stopping probability is

$$p^O_{\text{stop}}(s) \;=\; \frac{\exp\Bigl\{\beta_{\text{stop}} \cdot Q_\lambda(s,\perp)\Bigr\}}{\sum\limits_{a' \,\in\, \text{frontier}(s) \cup \{\perp\}} \exp\Bigl\{\beta_{\text{stop}} \cdot Q_\lambda(s,a')\Bigr\}}, \tag{7}$$

and when not stopping, the model expands node a with

$$p^{O}_{\text{select}}(s,a) \;=\; \frac{\exp\Bigl\{\beta_{\text{select}} \cdot Q_{\lambda}(s,a)\Bigr\}}{\sum\limits_{a' \,\in\, \text{frontier}(s)} \exp\Bigl\{\beta_{\text{select}} \cdot Q_{\lambda}(s,a')\Bigr\}}. \tag{8}$$

If $\beta_{\text{select}} = \beta_{\text{stop}}$, this reduces to a single softmax over the entire action space, but we allow them to differ for consistency with the heuristic models' flexibility.
The "myopic" variant replaces Q_{λ} with a *one-step approximate* value function $Q_{\lambda}^{\text{myopic}}$. For expansion actions, it assumes an immediate cost λ plus the expected value of stopping right after expansion:

$$Q_{\lambda}^{\text{myopic}}(s,a) \;=\; \mathbb{E}_{s' \sim T(\cdot|s,a)}\Bigl[r(s',\perp)\Bigr] \;-\; \lambda,$$

with $r(s',\perp)$ given by the reward function upon termination. This approximation can reduce computational demands yet still capture a tradeoff between continuing and stopping.

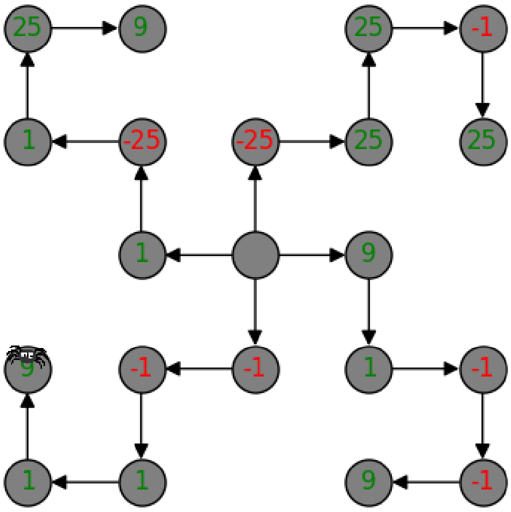
Experiment 2: Adapting to the Environment (Focusing on the Constant Condition)

Experiment 1 revealed that participants often relied on a best-first search strategy, which seemed particularly efficient given the reward distributions in that setting. However, the optimal model predicts that people should adapt their planning strategy to the structure of the environment rather than commit to a single approach. To test this prediction, Callaway et al. (2023) constructed three new environments for Experiment 2, each having the same underlying transition structure (four independent paths of five states each) but differing in how rewards are distributed. Specifically:

- **Constant variance**: All states share the same reward distribution (as in Experiment 1), thus making a best-first strategy well-suited.
- **Decreasing variance**: Extreme rewards only appear in the first state of each path, favoring a breadth-first approach to quickly discriminate promising versus unpromising paths.
- **Increasing variance**: Extreme rewards only occur in the last state of each path, making a depth-first strategy more effective at uncovering significant payoffs at the end of a branch.

By design, each of these distributions roughly favors a different classical search algorithm (breadth-first, best-first, or depth-first). Participants' performance in each environment indicated that they did indeed alter their planning behavior to better match the local reward structure. Figure 4 in the original study demonstrates that the model achieving the best effort-vs.-reward trade-off in a given environment also best fits the observed human data.

Our focus: Constant variance condition. In our extension, we specifically target the “constant variance” condition and further simplify the reward structure such that each node has the same probability of yielding a positive or negative payoff (e.g., ±25, ±9, ±1). The example in Figure 1 illustrates one such environment, with four paths extending from a central node. This setup effectively emulates the scenario in which every state is equally likely to contain high or low rewards, thus creating an environment where a best-first strategy appears intuitively sensible, yet leaves room for potential adaptations under different cognitive costs.



Figur 1: Illustration of the constant-variance environment used in our extension of Experiment 2. Each node conceals a reward (green or red), sampled from a common distribution (e.g. $\pm25, \pm9, \pm1$). This figure is a schematic; actual experiments randomize the location of each reward at the start of each trial.

As in Callaway et al. (2023), the rationale behind this manipulation is to examine whether human planners stick to a single search strategy (as one might expect from a purely best-first approach) or adapt to factors such as perceived risk, subjective cost, or time pressure. Our data collection modifies the original task to incorporate *cognitive control* variables (e.g., self-reported workload, time intervals between clicks) that might modulate planning depth and direction. In particular, we ask:

- **Adaptation of heuristics.** Do participants still gravitate toward a best-first approach in a constant reward setting, or can we observe variations when costs are high?
- **RL-based modeling.** Can a higher-level RL policy capture changes in the number of expansions, reflecting a real-time trade-off between exploration (revealing more nodes) and exploitation (terminating earlier to avoid cognitive costs)?
- **Outcome vs. effort.** How does the uniform risk across all states influence participants’ willingness to explore multiple paths? Do changes in subjective cost or time constraints lead to more rapid stopping?

Together, these questions extend the original Experiment 2 design to clarify how humans allocate cognitive resources even in “uniformly uncertain” environments. Our preliminary evidence suggests that although best-first expansions remain common, participants do *adapt* when certain cost parameters (e.g., mental effort, time penalty) become sufficiently salient. Thus, consistent with the broader findings of Experiment 2, people adjust not only to the external environment but also to their own internal constraints and preferences.

Control Deprivation Manipulation and the Rationale for an RL Approach

In addition to investigating resource-constrained planning, our study introduced an experimental manipulation designed to induce a sense of *control deprivation* among participants. Specifically, we adopted a common psychological procedure in which participants were asked to complete a series of computer-based quiz tasks of varying durations (e.g., a shorter vs. a longer time allotment). After submitting their responses, participants received feedback about the correctness of their answers. Crucially, in the “deprivation” condition, this feedback was intentionally *inaccurate*: although participants might have selected the correct response, we sometimes informed them that their response was wrong. Such false or discouraging feedback aims to diminish participants’ perceived efficacy and thus heighten their sense of frustration or lack of control.

Our motivation for incorporating this control-deprivation manipulation stems from the hypothesis that an individual’s feeling of control can significantly influence their decision-making processes, especially under constraints of time or cognitive resources. By varying both the time available (long vs. short quiz) and the accuracy of performance feedback, we sought to create a gradient of experienced control—ranging from a relatively empowered state to a distinctly deprived one. This gradient offers a window into how participants might shift their planning strategies and computational resource allocation when they perceive themselves as less or more in control of the situation.

Why Reinforcement Learning? We believe that modeling human decision-making in these scenarios using a *Reinforcement Learning* (RL) framework provides important benefits over purely normative or heuristic-based models. In particular:

- **Dynamic Adaptation:** RL models can track how participants adaptively tune their exploration and exploitation behaviors over multiple trials, especially under changing conditions of perceived control or feedback accuracy.
- **Cost–Benefit Analysis:** By framing each planning or search step as an action that incurs a cognitive or temporal cost, RL approaches naturally capture the trade-offs participants face. This aligns well with metalevel MDP perspectives, in which each computational step (e.g., expanding a node) is treated as a resource-intensive operation whose value must outweigh its cost.
- **Robustness to Noisy Feedback:** Because RL algorithms typically learn from observed rewards (and punishments) even if these signals are imperfect

or stochastic, they are well-suited to modeling conditions where participants receive misleading or manipulative feedback. The agent's policy evolves based on cumulative experience, mirroring the way real-world decision-makers gradually adjust strategy under uncertain or biased information.

Thus, by applying an RL-based modeling strategy to our control-deprivation design, we aim to gain deeper insights into how individuals reconfigure their planning or metacognitive policies in response to altered perceptions of control. In particular, we can assess whether participants' sense of learned helplessness or heightened vigilance (due to false feedback) leads them to reduce their search depth, shift to simpler heuristic expansions, or otherwise deviate from rational metareasoning predictions. Our preliminary findings suggest that RL models capture these adaptive shifts more accurately than static or purely normative accounts, reinforcing the view that human planning should be studied as an ongoing, feedback-driven process.

Methodology

Overview of the Extended Mouselab-MDP Task

We adopt a Mouselab-MDP style interface that externalizes planning operations by requiring participants to "click" on hidden rewards to reveal them. Under constant variance, each hidden reward R_i follows an identical distribution, but participants incur a small time or point cost for each click. Unlike the original design, we incorporate self-reported workload ratings (via brief questionnaires) to gauge individual differences in perceived planning cost.

Cognitive Control Modeling

We embed a *cognitive control* parameter α in the metalevel reward function to reflect each individual's sensitivity to planning costs. For instance, each node expansion might incur a cost $c(\alpha)$, which scales based on reported mental effort. This approach aligns with the notion that planning cost is subjective and context-dependent. We hypothesize that participants with higher α values exhibit shallower expansions (or switch to simpler heuristics earlier), whereas those with lower α might search more extensively. By fitting these models to participants' revealed sequences, we aim to test whether a single parameter can explain large inter-individual differences in planning strategies.

Reinforcement Learning Perspective

To tackle the combinatorial explosion of metalevel states, we use a reinforcement learning approach to approximate the optimal policy:

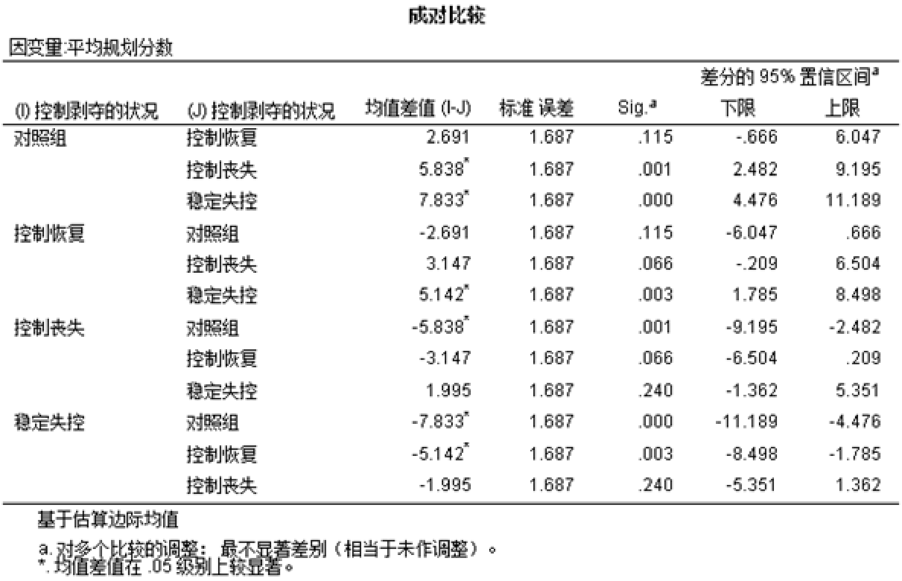
1. **State Representation:** Encodes which nodes are revealed, partial reward estimates, and an internal memory of expansions.
2. **Action Space:** Select a node to expand, or terminate and execute the best path found so far.

3. **Reward Function:** Reflects the net external reward minus cost of expansions, factoring in α to capture personal effort.

4. **Learning Algorithm:** A neural network approximator for $Q(\text{state}, \text{action})$ or a policy gradient method (e.g., PPO) to learn how to allocate expansions efficiently.

Results and Analysis

Impact of Control Deprivation on Planning Performance



成对比较

因变量:平均规划分数

					差分的 95% 置信区间[a]	
(I) 控制剥夺的状况	(J) 控制剥夺的状况	均值差值 (I-J)	标准 误差	Sig.[a]	下限	上限
对照组	控制恢复	2.691	1.687	.115	-.666	6.047
	控制丧失	5.838*	1.687	.001	2.482	9.195
	稳定失控	7.833*	1.687	.000	4.476	11.189
控制恢复	对照组	-2.691	1.687	.115	-6.047	.666
	控制丧失	3.147	1.687	.066	-.209	6.504
	稳定失控	5.142*	1.687	.003	1.785	8.498
控制丧失	对照组	-5.838*	1.687	.001	-9.195	-2.482
	控制恢复	-3.147	1.687	.066	-6.504	.209
	稳定失控	1.995	1.687	.240	-1.362	5.351
稳定失控	对照组	-7.833*	1.687	.000	-11.189	-4.476
	控制恢复	-5.142*	1.687	.003	-8.498	-1.785
	控制丧失	-1.995	1.687	.240	-5.351	1.362

基于估算边际均值

a. 对多个比较的调整：最不显著差别（相当于未作调整）。

*. 均值差值在 .05 级别上较显著。

Figur 2: Post-hoc pairwise comparisons of mean planning scores

We conducted an analysis of variance (ANOVA) with the control deprivation condition as a between-subjects factor. The results indicate a significant main effect of control deprivation, $F = -8.347$, $p = 0.000$, $\eta_p^2 = 0.239$. No significant interaction effect was found between the *stability* of deprivation and the *duration* of deprivation, suggesting that participants' performance was primarily influenced by the overall status of their perceived control rather than how long or how stably it was removed.

Post-hoc pairwise comparisons. Using post-hoc tests to compare each condition:

- **Control vs. Stable Control Deprivation.** The control group performed significantly better than the stable deprivation group ($t = 7.833, p = 0.000$, 95% CI $= [4.476, 11.189]$).

- **Control vs. Loss of Control.** The control group also outperformed the loss-of-control group ($t = 5.838, p = 0.001$, 95% CI $= [2.482, 9.195]$).

- **Recovery vs. Stable Control Deprivation.** Participants who experienced recovery performed significantly better than those with stable deprivation ($t = 5.142, p = 0.003$, 95% CI = $[1.785, 8.498]$).

From these results, it is clear that removing or reducing participants' sense of control leads to reliably lower planning scores relative to the control group. Notably, however, the analyses revealed no significant difference in another key measure (***), implying that the basic decision-making process might be fundamentally similar across groups. Consequently, we can apply the same RL-based model to each subset of participants' data, allowing parameter estimates (e.g., cost or exploration parameters) to capture individual or subgroup differences without needing entirely separate model structures.

We evaluated four different models—Breadth, Depth, Best, and Optimal—alongside the aggregate Human data, focusing on how well each model fits participants' decisions under our RL-based framework. Below, we present both quantitative (*NLL*, *AIC*, and *GML*) and qualitative (behavioral curves) comparisons.

Model Fit Using Geometric Mean Likelihood (GML)

To gauge how closely each model's policy predictions align with observed choices, we computed the negative log-likelihood (NLL) for each trial and then derived the *geometric mean likelihood* (GML). For a total of N_ACT actions (clicks) and a sum of test negative log-likelihood $\sum$ Test_NLL, the average $\overline{\text{NLL}}$ is:

$$\overline{\text{NLL}} = \frac{\sum \text{Test\_NLL}}{\text{N\_ACT}},$$

and we define

$$\text{GML} = \exp\Big(-\overline{\text{NLL}}\Big).$$

A higher GML indicates that the model yields higher likelihood for the human data.

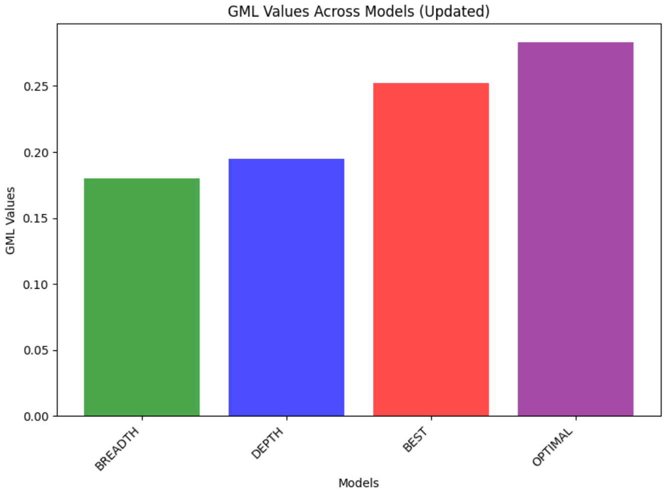
Model	$\sum$ Test_NLL	$\overline{\text{NLL}}$	GML
Breadth	16760.16	1.713	0.180
Depth	16012.89	1.637	0.195
Best	13492.14	1.379	0.252
Optimal	12347.41	1.262	0.283

Tabel 1: Geometric Mean Likelihood (GML) for Different Models. Higher GML suggests a better fit to human data.

Figure 3 visualizes these GML values. We see that Optimal attains the highest score, followed by Best, Depth, and Breadth.

Comparison of Strategy Curves

We next compared how each model's policy accumulates reward as the number of expansions (clicks) increases, and how it adjusts the second-click decision based on the first revealed value. Figures 4 and 5 illustrate these trends:



Figur 3: GML values across models. Optimal obtains the highest geometric mean likelihood, indicating it better captures participants' observed decisions compared to the other strategies.

1. **Expected Reward vs. Number of Clicks** (Figure 4): Optimal and Best generally track human performance more closely over a range of click counts, whereas Breadth lags behind until many nodes are revealed. Depth initially escalates more quickly than Breadth but still fails to match the final performance of Best and Optimal.

2. **Proportion of Second Clicks on the Same Path** (Figure 5): We grouped trials by the value revealed on the first click and measured how often a second click remains on the same path. A strong dependence on the revealed value indicates a *best-first* tendency, while a uniform pattern suggests *breadth-first* or *depth-first* expansions. In the human data (black dots), we observe a steep increase in same-path exploration for higher initial values, echoing Best and Optimal more closely than Breadth or Depth.

AIC Computation and Model Ranking

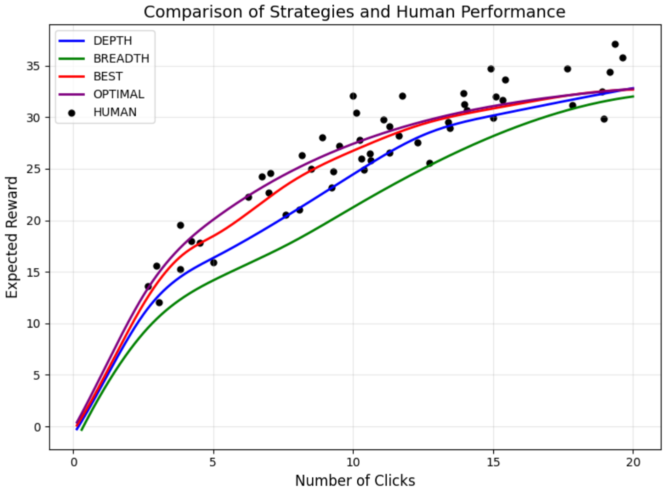
To further distinguish among models, we computed Akaike's Information Criterion (AIC), which penalizes model complexity in addition to raw likelihood. Let θ denote a model's parameter vector of dimension k. For data $\{(x_i, y_i)\}_{i=1}^n$, we define

$$\mathrm{NLL}(\theta) = -\sum_{i=1}^{n} \ln\Big[f\Big(y_i \mid x_i,\, \theta\Big)\Big].$$

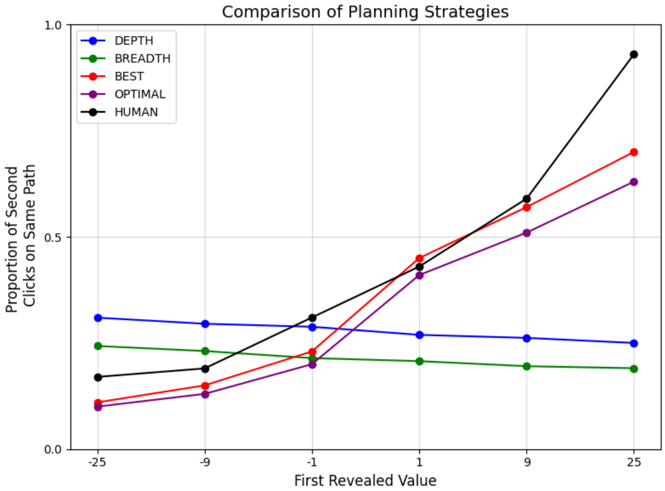
The AIC is then

$$\mathrm{AIC} \;=\; 2\,k + 2\,\mathrm{NLL}(\hat{\theta}),$$

where $\hat{\theta}$ is the MLE parameter estimate. Table 2 shows the AIC scores and the resulting Akaike weights. Optimal emerges as the best-fitting model according to AIC, with a much lower score relative to the other approaches.



Figur 4: Comparison of Strategies and Human Performance. Each curve shows expected reward as a function of how many nodes the agent expands (clicks). Black dots indicate human participants' aggregated data.



Figur 5: Proportion of second-click expansions on the same path as a function of the first revealed node's value. Human data (black) reveals a strong "best-first" tendency for high positive values, closely resembling the Optimal and Best curves.

Discussion of Findings

Collectively, these results suggest that while heuristic models (Breadth, Depth, Best) can approximate partial aspects of human behavior, the Optimal model provides a substantially better fit overall. This advantage is evident in terms of negative log-likelihood (reflected in higher GML) and penalized likelihood measures (AIC). Qualitative comparisons of how quickly each model escalates reward and how it reacts to initial high or low values further confirm that people's strategy is closer to Optimal under this RL-based framework, at least within the "constant variance" environment we tested.

Moreover, the strong performance of Optimal highlights that human participants appear sensitive to both reward magnitudes and the cost of information. When

Model	k	$\sum$ Test_NLL	AIC	Δ_i	$\exp(-\Delta_i/2)$	Akaike Weight
Breadth	3	16760.16	33526.32	8823.50	≈ 0.0	≈ 0.0
Depth	3	16012.89	32031.78	7328.96	≈ 0.0	≈ 0.0
Best	3	13492.14	26990.28	2287.46	≈ 0.0	≈ 0.0
Optimal	4	12347.41	24702.82	0	1.0	1.0

Tabel 2: AIC Computation and Akaike Weights. Optimal dominates in terms of both AIC and associated weight.

large negative or positive outcomes are revealed, participants are more likely to adopt or abandon that path accordingly, mirroring the best-first component within Optimal.

In summary, consistent with [1], human planning in this environment shows adaptive patterns that neither purely match a single heuristic nor deviate so randomly as to suggest no systematic structure. Our extensions further indicate that an RL-based perspective, where planning itself is treated as a sequential decision problem, can help unify these results by modeling how participants dynamically balance exploration cost and potential reward.

Discussion and Ongoing Work

By extending the constant variance condition of Callaway et al. (2023), we provide new evidence that individuals' planning processes exhibit varying degrees of *cognitive control* and can be partially captured by a metalevel MDP-based RL framework. This aligns with prior claims that planning is itself a sequential decision problem, with computations as actions and belief states as internal representations. Our approach further suggests that human resource allocation in planning can be systematically studied through the lens of RL, bridging cognitive psychology and AI research. One open question is how best to incorporate hierarchical RL for more complex tasks, in which subproblems can themselves be decomposed into smaller metalevel tasks. Another direction is to incorporate real-time psychophysiological or neural measures to refine α estimates and further validate the notion of cognitively controlled expansions.

Conclusion

We have introduced an extended metalevel MDP model of resource-constrained planning, grounded in the "constant variance" setting of Experiment 2 in [1] and enriched by a cognitive control component. Our collaboration with psychology researchers allowed us to integrate subjective workload into the metalevel cost function, offering a more nuanced view of how people decide when and where to invest mental effort. Additionally, we demonstrated how Reinforcement Learning techniques can be employed to approximate the metalevel policy for planning, emphasizing that planning itself is a sequential decision-making process at a higher level. Future work includes validating our cognitively controlled model across broader task domains, scaling up the RL framework to handle even more complex or high-dimensional planning problems, and exploring richer forms of individual differences in metacognitive awareness. We hope this research helps bridge the gap between cognitive modeling and practical AI systems that must allocate their limited computational resources adaptively.

Litteratur

[1] F. Callaway, B. van Opheusdena, S. Gulb, P. Dasc, P. Kruegerd, T. L. Griffithsa,d, and F. Liedere, “Rational use of cognitive resources in human planning,” *Proceedings of the National Academy of Sciences*, 2023.

[2] T. L. Griffiths, F. Lieder, and N. D. Goodman, “Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic,” *Topics in Cognitive Science*, vol. 7, no. 2, pp. 217–229, 2015.

[3] J. W. Payne, J. R. Bettman, and E. J. Johnson, *The Adaptive Decision Maker*, Cambridge University Press, 1993.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 2nd edition, 2018.