What are the mechanisms underlying metacognitive learning in the context of planning?

Ruiqi He (ruiqi.he@tuebingen.mpg.de)
Max Planck Institute for Intelligent Systems
Stuttgart, Germany

Falk Lieder (falk.lieder@tuebingen.mpg.de)
Max Planck Institute for Intelligent Systems
Stuttgart, Germany

Abstract

How is it that humans can solve complex planning tasks so efficiently despite limited cognitive resources? One reason is its ability to know how to use its limited computational resources to make clever choices. We postulate that people learn this ability from trial and error (*metacognitive reinforcement learning*). In this work, we systematize models of the underlying learning mechanisms and enhance them with more sophisticated additional mechanisms. We fit the resulting 86 models to human data collected in previous experiments where different phenomena of metacognitive learning were demonstrated and performed Bayesian model selection. Our results suggest that a gradient ascent through the space of cognitive strategies can explain most of the observed qualitative phenomena, and is, therefore, a promising candidate for explaining the mechanism underlying metacognitive learning.

Keywords: metacognitive learning, planning, strategy discovery, cognitive modelling, reinforcement learning

Introduction

Humans frequently face complex problems that require planning long chains of actions to accomplish far-off objectives. A search tree can represent the space of potential future actions and outcomes, which expands exponentially as the length of the sequences increases. While exponential growth in computational power enables current trends in artificial intelligence, the cognitive capabilities of the human mind are much more constrained. So, how is it possible that people can still plan so efficiently? One potential explanation is that meta-reasoning, the ability to reason about reasoning, might help people to accomplish more with less computational effort (Griffiths et al., 2019). In the context of planning, this means making wise choices about when and how to plan, that is whether and how to efficiently make use of limited cognitive resources (*resource-rationality*) (Lieder & Griffiths, 2020). However, according to Russell and Wefald (1991), optimal meta-reasoning is often regarded as an intractable problem. This raises the question of how people can nonetheless solve the intractable meta-reasoning problem. One possibility is that people learn an approximate solution via trial and error, an idea known as *metacognitive reinforcement learning* (Lieder & Griffiths, 2017; Krueger, Lieder, & Griffiths, 2017; Lieder, Shenhav, Musslick, & Griffiths, 2018). This idea has been used in earlier research to explain how people learn to select between various cognitive strategies (Erev & Barron, 2005; Rieskamp & Otto, 2006; Lieder & Griffiths, 2017), how many steps to plan ahead (Krueger et al., 2017) and when to exercise how much cognitive control (Lieder et al., 2018). In the context of planning, previous work suggests that metacognitive reinforcement learning adapts people's planning strategies to their environments (Jain, Callaway, & Lieder, 2019; He, Jain, & Lieder, 2021b) and adapts how much planning they perform (He, Jain, & Lieder, 2021a). While previous work each focused on explaining individual aspects of metacognitive learning with a small set of models, none of the models was tested to explain both aforementioned observed qualitative phenomena in the context of planning. In addition, previous findings paint a rather inconsistent and even contradictory picture of how people learn planning strategies, with different articles arguing for different learning mechanisms (Jain, Gupta, et al., 2019; He et al., 2021a).

Therefore, in this work, we investigate whether there is one metacognitive reinforcement learning model that can largely explain both observed phenomena of adaptation to environment structures and adapting the amount of planning to its cost and benefits. Our contribution is two-fold: i) We systematically compare existing models on data collected in empirical experiments, and ii) we extend existing models to systematically formalize plausible alternative assumptions and all of their possible combinations. This led to 86 different models, which we fit using maximum likelihood criterion and compare using Bayesian model selection, as well as perform model simulation. The winning model gives us an indication of the underlying mechanisms of how people learn planning strategies.

This line of research contributes to the larger goal of understanding metacognitive learning. It also provides a foundation for training programs aiming to improve human decision-making and to help people overcome maladaptive ways of learning planning strategies.

Background

To model the mechanism of metacognitive learning, we take inspiration from reinforcement learning algorithms and use the framework of meta-decision-making, which we will now briefly introduce and explain how they can be combined into a framework called *metacognitive reinforcement learning*.

Reinforcement learning

Previous studies suggest that human learning is motivated by reward and penalties gained through trial and error (Niv,

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

2009), which builds the foundation of reinforcement learning algorithms that learn to predict the potential reward from performing a specific action a in a specific state s. This estimate $Q(s,a)$ is updated according to the reward prediction error δ, which is the difference between actual and expected rewards:

$$Q(s,a) \leftarrow Q(s,a) - \alpha \cdot \delta \tag{1}$$

where Q denotes the Q-value (Watkins & Dayan, 1992) and α is the learning rate.

Meta-decision-making

The brain is supposedly equipped with multiple decision systems that interact in various ways (Dolan & Dayan, 2013; Daw, 2018). The model-based system, in contrast to Pavlovian and model-free systems, allows for flexible reasoning about which action is preferable but demands a process for deciding which information should be considered for a given decision. Therefore, an important part of deciding how to decide is to efficiently balance decision quality and decision time, known as *meta-decision-making* (Boureau, Sokol-Hessner, & Daw, 2015). The problem of meta-decision-making has been recently formalized as a meta-level MDP (Krueger et al., 2017; Griffiths et al., 2019):

$$M_{meta} = (\mathcal{B}, \mathcal{C} \cup \{\perp\}, T_{meta}, r_{meta}), \tag{2}$$

where belief states $b_t \in \mathcal{B}$ denotes the model-based decision system's beliefs about the values of actions. The computations of the decision system $(c_1, c_2, \cdots)$ probabilistically determine the temporal development of those belief states $b_1, b_2, \cdots$ according to the meta-level transition probabilities $T_{\text{meta}}(b_t, c_t, b_{t+1})$. The meta-level reward function $r_{\text{meta}}(b_t, c_t)$ encodes the cost of performing the planning operation $c_t \in \mathcal{C}$ and the expected return of terminating planning ($c_t = \perp$) and acting based on the current belief state b_t. Reinforcement learning algorithms, such as Q-learning (see Equation 1), can be used to solve this meta-level MDP.

Metacognitive reinforcement learning

Finding efficient planning strategies can be formalized as solving a meta-level MDP for the best meta-level policy (Griffiths et al., 2019). However, as it is often computationally intractable to solve meta-decision-making problems optimally, we will assume that the brain approximates optimal meta-decision-making through reinforcement learning mechanisms (Russell & Wefald, 1991; Callaway, Gul, Krueger, Griffiths, & Lieder, 2018) that attempt to approximate the optimal solution of the meta-level MDP defined in Equation 2 by either learning to approximate the optimal policy directly (He et al., 2021b) or by learning an approximation to its value function (Jain, Callaway, & Lieder, 2019).

Experiments

To test the ability of our models to explain different aspects of metacognitive learning, we used data from previous work that examined several aspects of it in the domain of planning.

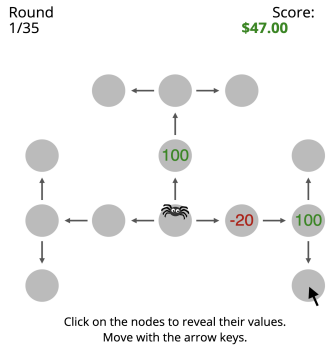


Figure 1: Exemplary trial of the planning task

He et al. (2021b) and He et al. (2021a) recruited 382 participants on CloudResearch and utilized the Mouselab-MDP paradigm (Callaway, Lieder, Krueger, & Griffiths, 2017) to design two experiments where participants were asked to perform 35 repeated trials of a planning task (see Figure 1). The goal in the experiment was to collect a high score, which signals the adaptiveness and resource-rationality (Lieder & Griffiths, 2020) of the participant at a given trial. The rewards are initially hidden but can be revealed by clicking on the nodes. Each click has a cost. Participants' clicks were recorded as they indicate planning operations that people perform to estimate the values of alternative future locations.

Adaptation to different environment structures In the first experiment, 174 participants were evenly randomly allocated to one of three conditions, where the environment structure rendered either long-term planning (examining the farthest nodes), short-term planning (examining immediate nodes) or best-first search planning (starting with examining immediate and middle nodes and continue to examine other nodes according to the most promising ones) most beneficial. Analysis based on the collected click sequences suggested that people gradually learn to use the corresponding adaptive strategies for each environment.

Adaptation of the amount of planning depending on the costs and benefits of planning The second experiment indicated that people do learn how much to plan. For this, 208 participants were assigned to one of four different conditions, each differed in the benefit (high vs. low) and cost (high vs. low) of planning. Their number of clicks indicated whether participants learned to adapt their amount of planning depending on the condition.

Models and methods

The models of metacognitive learning we test in this article have three components: i) the representation of the planning strategies that the learning mechanism operates on, ii) the basic learning mechanisms, and iii) additional attributes. The following three sections introduce these components as well as describe how the models were fit and selected.

Mental representation of planning strategies

The planning strategies were modelled as softmax policies that depend on a weighted combination of 56 features (Jain et al., 2022). For instance, one group of features was related to pruning (Huys et al., 2012), which was associated with giving a negative value to consider a path whose predicted value was below a specific threshold. Therefore, using this representation, a person's learning trajectory can be described as a time series of the weight vectors that correspond to their planning strategies in terms of those features.

Basic learning mechanisms

We considered three possible basic learning mechanisms: learning the value of computation, gradient ascent through the strategy space, forming a mental habit, and no learning. All learning mechanisms approximate the meta-level Q-value by a linear combination of the features mentioned above and a set of learned weights:

$$Q_{\text{meta}}(b_k, c_k) \approx \sum_{j=1}^{56} w_j \cdot f_j(b_k, c_k), \tag{3}$$

To compromise between exploitation and exploration, the actions are chosen probabilistically, maximising the predicted action value, by using the softmax rule (Williams, 1992) $P(c_k|b_k, Q_{\text{meta}}) \propto \exp(Q_{\text{meta}}(b_k, c_k)/\tau)$ where τ is the inverse temperature parameter.

Learning the value of computation According to the Learned Value of Computation (LVOC) model, people learn how valuable it is to perform each planning operation depending on what is already known (Krueger et al., 2017). That is people discover and change their strategy continuously by learning to predict the values of planning operations. The weights in Equation 3 are learned by Bayesian linear regression of the bootstrap estimate $\hat{Q}(b_k, c_k) = r_{\text{meta}}(b_k, c_k) + \sum_{j=1}^{56} \mu_{j,k} \cdot f_j(b_{k+1}, c_{k+1})$ which is the sum of the immediate meta-level reward and the anticipated value of the future belief state b_{k+1} under the present meta-level policy. The predicted value of b_{k+1} is the scalar product of the posterior mean μ_t of the weights $\mathbf{w}$ given the observations from all preceding planning operations and the features $\mathbf{f}(b_{k+1}, c_{k+1})$ of b_{k+1} and the cognitive operation c_{k+1} that the current policy picks given state. To make the k^{th} planning operation, n weight vectors are sampled from the posterior distribution P using a generalized Thompson sampling $\tilde{w}_k^{(1)}, \cdots, \tilde{w}_k^{(n)} \sim P(\mathbf{w}|\mathcal{E}_k)$, where the set $\mathcal{E}_k = \{e_1, \cdots, e_k\}$ contains the meta-decision-maker's experience from the first k meta-decisions. Each meta-level experience $e_i \in \mathcal{E}_k$ is a tuple $\left(b_i, h_i, \hat{Q}(b_i, c_i; \mu_i)\right)$ containing a meta-level state, the selected planning operation in it, and the bootstrap estimates of its Q-value. The arithmetic mean of the sampled n weight vectors is then used to predict the Q-values of each potential planning operation $c \in \mathcal{C}$ according to Equation 3. The LVOC model therefore has the following free parameters: p, the mean vector μ_{prior} and variance σ^2_{prior} of its prior distribution $\mathcal{N}(\mathbf{w}; \mu_{\text{prior}}, \sigma^2_{\text{prior}} \cdot \mathrm{I})$ on the weights $\mathbf{w}$, the number of samples n and the inverse temperate τ.

Gradient ascent through the strategy space According to the REINFORCE model (Jain, Callaway, & Lieder, 2019; Williams, 1992), people adjust their planning strategy directly using gradient ascent through the space of possible planning strategies. When a plan is executed according to policy $\pi_{\mathbf{w}}$ and its outcomes are observed, the weights w representing the strategy are adjusted in the direction of the gradient of the return, which is the sum of the rewards on the chosen path minus the cost of the performed planning operations:

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha \cdot \sum_{t=1}^{O} \gamma^{t-1} \cdot r_{meta}(b_t, c_t) \cdot \nabla_{\mathbf{w}} \ln \pi_{\mathbf{w}}(c_t|b_t), \tag{4}$$

γ is the discount factor, and O is the number of planning operations executed by the model on that trial. The learning rate α is optimised using state-of-the-art optimizer ADAM (Kingma & Ba, 2014). The model has three free parameters: α, γ and τ that are fit separately for each participant. The weights are initialised randomly.

Mental habit formation This model assumes that the only mechanism through which people's planning strategies change is the formation of mental habits. Following Miller, Shenhav, and Ludvig (2019) and Morris (2022), this model assumes that people's propensity to perform a (type of) planning operation increases with the number of times they have performed it in the past. This is implemented as a softmax decision rule applied to a weighted sum of frequency-based features, including the number of previous clicks on the same node, the same branch, and the same level, respectively.

Non-learning model This model does not perform any parameter updates and does not use habitual features.

Extensions

We augmented the REINFORCE and LVOC models with three optional components: a two-stage hierarchical meta-decision-making process (*hierarchical meta-control*), metacognitive rewards for generating valuable information (*pseudo-rewards*), and deliberating about the value of termination when taking an action (*termination deliberation*).

Hierarchical meta-control Previous research suggests that foraging decisions are made by two distinct decision systems: the ventromedial prefrontal cortex and the dorsal anterior cingulate cortex (Rushworth, Kolling, Sallet, & Mars, 2012). We, therefore, developed an extension that first decides whether to continue planning (Stage 1) and if yes, selects the next planning operation according to either the LVOC or the REINFORCE model (Stage 2). For Stage 1, our models consider three potential decision rules. Each decision rule is a tempered sigmoid function $\sigma(x, \tau) = (1 + e^{-\frac{x}{\tau}})^{-1}$ (Papernot, Thakurta, Song, Chien, & Erlingsson, 2021). In each case, the function's argument x is a different function $f(\mathbb{M})$ of the

expected sum of rewards along the best path according to the information observed so far ($\mathbb{M} = \max_{path} \mathbb{E}[R(\text{path}) \mid b]$). Concretely, the three stopping rules compare $\mathbb{M}$ against a fixed threshold, a threshold that tracks the outcomes of previous trials, and a threshold that decreases with the number of clicks, respectively.

Fixed threshold This decision rule probabilistically terminates planning when the normalized value of $\mathbb{M}$ reaches the threshold η, that is $P(C = \perp \mid b) = \sigma\left(\frac{\mathbb{M} - v_{\min}}{v_{\max} - v_{\min}} - \eta, \tau\right)$, where $v_{\min}$ and $v_{\max}$ are the trial's lowest and highest possible returns, respectively.

Decreasing threshold Building on the observation that the threshold of the resource-rational planning strategy decreases with the number of clicks (Callaway, Lieder, et al., 2018), this decision rule adjusts the threshold based on the number of clicks made so far (n_c), that is: $P(C = \perp \mid b) = \sigma(\mathbb{M} - e^a + e^b \cdot n_c, \tau)$, where a and b are the free parameters.

Threshold based on past performance This decision rule models the idea that people learn what is good enough from experience. Concretely, this decision rule assumes that the threshold $M \sim \mathcal{N}\left(m; \frac{\eta}{\sqrt{n+1}}\right)$ is a noisy estimate of the average m of their previous scores, that is $P(C = \perp \mid b) = \sigma(\mathbb{M} - M), \tau)$, where n is the number of trials and η is a free parameter. The probability distribution of the threshold is derived from the assumption that the threshold is an average of noisy memories of previous scores.

Pseudo-rewards The central role of reward prediction errors in reinforcement learning (Schultz, Dayan, & Montague, 1997; Glimcher, 2011) and the dearth of external reward in metacognitive learning (Hay, 2016) indicate that the brain might accelerate the learning process by producing additional metacognitive pseudo-rewards that convey the value of the information produced by the last planning operation. Concretely, the pseudo-reward (PR) for transitioning from belief state b_t to b_{t+1} is the difference between the expected value of the path that the agent would have taken in the previous belief state b_t and the expected value of the best path in the new belief state b_{t+1}: $\text{PR}(b_t, c, b_{t+1}) = \mathbb{E}[R_{\pi_{b_{t+1}}} | b_{t+1}] - \mathbb{E}[R_{\pi_{b_t}} | b_{t+1}]$ where $\pi_b(s) = \text{argmax}_a \mathbb{E}_b[R \mid s, a]$ is the policy the agent will use to navigate the physical environment when its belief state is b, and R is the expected value of the sum of the external rewards (e.g., the sum of rewards collected by moving through the planning task) according to the probability distribution b.

Termination deliberation If people engaged in rational metareasoning (Griffiths et al., 2019), they would calculate the expected value of acting on their current belief b from the information it encodes (*termination deliberation*). Alternatively, people might learn when to terminate through the same learning mechanism through which they learn to select between alternative planning operations (no termination deliberation).

Model fitting

Combining the basic learning mechanisms with the model attributes resulted in 86 different models. We fitted all models to 382 participants from both experiments by maximizing the likelihood function of the participants' click sequences using 400 iterations of Bayesian optimization (Bergstra, Yamins, & Cox, 2013). The likelihood of a click sequence is the product of the likelihood of the individual clicks.

Model selection

To select the model that best explains the observed behavior, we estimated the expected proportion of people who are best described by a given model (r) and the exceedance probability ϕ that this proportion is significantly higher than the corresponding proportion for any other model by using random effect Bayesian model selection (BMS) (Rigoux, Stephan, Friston, & Daunizeau, 2014). To obtain the equivalent conclusions for groups of models that share some feature, we performed family-level Bayesian model selection (Penny et al., 2010). To ensure robustness and reproducibility, we used bootstrapping (Wehrens, Putter, & Buydens, 2000), that is, we fitted the models twice and used the results to generate 1000 synthetic data sets. The BMS results were averaged across all bootstrap samples.

Results

An overview of the 86 models and corresponding features as well as the code can be found in `https://osf.io/wz9uj/`.

Comparing all models for all participants

To examine which of the learning mechanisms can best explain human behavior, we grouped the models into 4 model families: non-learning, mental habit, LVOC, and REINFORCE models. We found that the model family whose members provided the best explanation for the largest number of participants was REINFORCE (see Table 1), which explained about 43.82% of the participants better than models from other model families. The second most successful model family was the non-learning model. It provided the best explanation for 30.87% of the participants, which was mainly driven by the high proportion of participants who did not show any signs of learning. Therefore, to examine the actual learning behavior, for the remaining analysis, we focused on participants who demonstrated learning. That is, participants who changed their planning strategies at least once in the first experiment and participants whose planning amount changed significantly during the second exper-

Table 1: Family-level BMS for learners

Model family	r	ϕ
Non-learning	0.31	0.01
Mental habit	0.07	0
LVOC	0.18	0
REINFORCE	0.43	0.99

Table 2: Family-level BMS for learners

Model family	r	ϕ
Non-learning	0.25	0
Mental habit	0.07	0
LVOC	0.21	0
REINFORCE	0.47	1

iment (determined using the Mann-Kendall test of trend; all $S > 105$ for increasing trend, all $S < -57$ for decreasing trend, all $p < .05$). This led to the selection of 224 participants (58.64%).

Comparing all models for learners

Family-level BMS for the remaining participants showed a decrease in the proportion of participants best explained by the non-learning model to 24.92%. REINFORCE models now explained the data from 47.41% of the learners better than the other models (see Table 2).

Table 3: Model-level BMS of the learning models that achieved $r > 0.05$ for all learners across both experiments.

Model	r	ϕ
Plain REINFORCE	0.10	0.79
REINFORCE with PR	0.08	0.10
REINFORCE with TD	0.05	0.03
REINFORCE with PR and TD	0.05	0.01
Plain LVOC	0.06	0.02

Comparing the learning models individually across both experiments, plain REINFORCE provided the best explanation for the highest proportion of participants, followed by its variants and the plain LVOC model (see Table 3).

To shed more light on whether the additional model attributes contribute towards explaining human metacognitive learning, we conducted BMS family-level comparing models with and without the model attributes across both experiments. The results suggested that more than half of the participants are better explained by models without pseudo-reward ($r_{\text{no PR}} = 0.61, \phi_{\text{no PR}} = 0.99$), without termination deliberation ($r_{\text{no TD}} = 0.61, \phi_{\text{no TD}} = 0.99$) and without hierarchical meta-control ($r_{\text{no HR}} = 0.75, \phi_{\text{no HR}} = 1, r_{HR} = 0.25, \phi_{HR} = 0$). The plain REINFORCE learning mechanism explained the largest proportion of participants (see Table 4 and 5), again followed by its variants in Experiment 1 and variants of LVOC with pseudo-reward and termination deliberation in Experiment 2 (see Table 6).

Yet, almost 40% of participants did appear to leverage pseudo-rewards ($r_{\text{PR}} = 0.39, \phi_{\text{PR}} = 0.01$) and termination deliberation ($r_{\text{TD}} = 0.39, \phi_{\text{TD}} = 0.01$), respectively. Comparing the difference in Bayesian Information Criterion (Schwarz, 1978) between the REINFORCE model with pseudo-rewards and its plain version for all learners revealed substantial evidence for its presence in 106 out of 224 participants and substantial evidence for its absence in 118 other participants. Conducting the same analysis for termination deliberation resulted in 87 participants in favor of it, while 137 were better fitted by plain REINFORCE. A χ^2 test comparing the proportion of participants whose data is better explained by a model with pseudo-rewards (42% vs. 37%, $\chi^2(3) = 0.73, p = .86$) and with termination deliberation (35% vs. 31%, $\chi^2(3) = 3.02, p = .39$) between the two experiments yielded no significant differences. Therefore, the difference between the learning behavior of people regarding pseudo-reward and termination deliberation could not be explained by situational factors, but might rather be due to inter-individual differences.

Table 4: Family-level BMS for Experiment 1

Model family	r	ϕ
LVOC	0.19	0
REINFORCE	0.81	1

Table 5: Family-level BMS for Experiment 2

Model family	r	ϕ
LVOC	0.41	0.09
REINFORCE	0.59	0.91

Table 6: Model-level BMS of the models that achieved $r > 0.05$ for Experiments 1 and 2

Exp	Model	r	ϕ
1	Plain REINFORCE	0.13	0.56
1	REINFORCE with PR	0.10	0.07
1	REINFORCE with TD	0.07	0.04
2	Plain REINFORCE	0.10	0.35
2	REINFORCE with PR	0.08	0.09
2	LVOC with TD	0.06	0.17
2	LVOC with PR	0.06	0.16

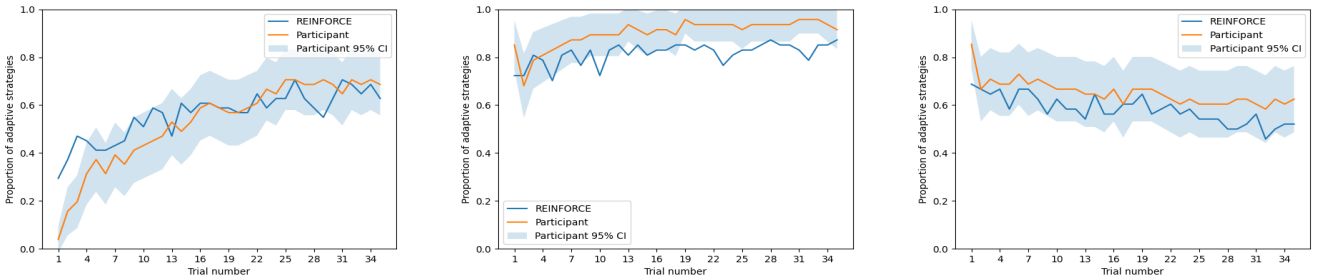
Robustness The model selection results were highly robust. For all reported r-values, the width of the 95% confidence intervals was at most ± 0.01. At the family-level, REINFORCE models provided the best explanation for the largest proportion of participants in 99% of all bootstrap samples, and the plain REINFORCE was the best individual model in 95% of all bootstrap samples.

How well can our best models capture the qualitative changes in people's planning strategies?

While our analysis indicated the existence of inter-individual differences regarding the additional model enhancements, there was agreement on REINFORCE being the most promising basic learning mechanism. Therefore, to examine, how well plain REINFORCE can explain all phenomena observed in both experiments, we simulated participants' behavior in the three conditions of Experiment 1 and the two conditions of Experiment 2 with the fitted model parameters[1]. Figure 3 shows the increasing trend in the predicted level of resource-rationality over time across both experiments (Mann-Kendall test: all $S > 245$ and $p < .01$ for both models and participants). This shows that REINFORCE could capture the observed increase in adaptiveness.

Figure 2 displays the proportion of adaptive planning strategies in the first experiment. To determine whether a participant used an adaptive planning strategy on a given trial, we inspected the first click in each trial, which signals what kind of strategy has been used. First click on the farthest node signals the adaptive far-sighted strategy in the first condition; first click on an immediate node signals the near-sighted strategy in the second condition, and first click on the immediate and middle nodes signals the best-first-search in the third condition. REINFORCE captured that people learned to in-

[1]For the visualizations, we refitted the model with a higher number of optimization iterations as it affected the model performance.



(a) Far-sighted planning environment (b) Short-sighted planning environment (c) Best-first-search planning environment

Figure 2: Comparing average proportion of adaptive planning strategies between the participants and the fitted model inferred by proportion of click sequences consistent with the corresponding optimal strategies in Experiment 1

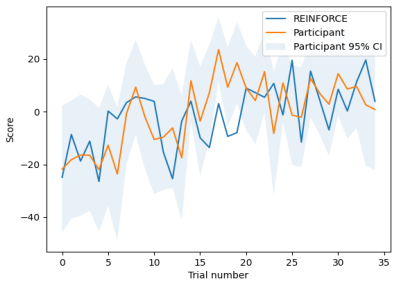
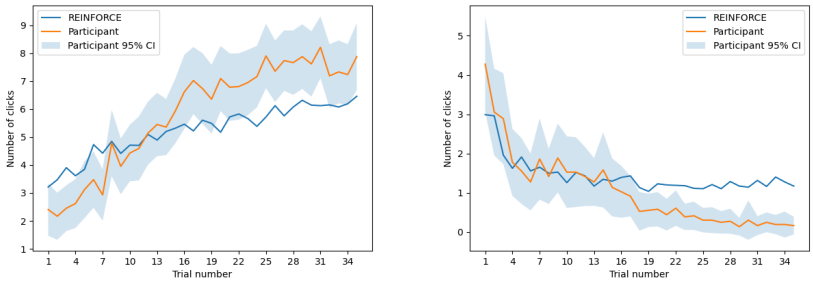


Figure 3: Average score of the fitted model and of the participants across 35 trials for both experiments



(a) Planning is beneficial (b) Planning is not beneficial

Figure 4: Average number of clicks of the fitted model and of the participants across 35 trials for Experiment 2

creasingly more often rely on adaptive strategies in the condition, where far-sighted planning is beneficial (see Figure 2a, Mann-Kendall test: increasing proportion of adaptive strategies for both the model and participants; all $S > 395, p < .01$) as well as when the environment favored near-sighted planning (see Figure 2b; increasing trend for participants and REINFORCE: $S > 263, p < .01$). Moreover, the model captured that participants appeared to use increasingly fewer adaptive strategies in the environment that preferred best-first-search planning (see Figure 2c; $S < -377; p < .01$).[2]

REINFORCE also partly captured the participants' learning behavior in Experiment 2. For the conditions where planning was beneficial, the model correctly predicted that the amount of planning would increase significantly over time (see Figure 3a; Mann-Kendall test: $S = 506, p < .01$). For the condition where planning is less beneficial, the model predicted that the number of clicks would decrease to a nearly optimal level (see Figure 3b, Mann-Kendall test: $S = -332, p < .01$). Participants learned to decrease their amount of planning to an even greater extent and converged on planning less than the resource-rational strategy. This indicates that participants experience an additional cost that is not yet captured by our models.

[2]This might reflect shortcomings of the rule He, Jain, & Lieder (2021) used to classify people's strategies in this environment.

Discussion and further work

In this article, we tested 86 computational models of how people learn planning strategies against data collected in two experiments that tested different characteristics of metacognitive learning, namely the adaptation to different environment structures and the adaptation to different levels of planning costs based on the proportion of adaptive strategies, the achieved score and the amount of planning. Overall, we found consistent evidence that the learning mechanism REINFORCE can largely capture the observed phenomena, like learning far-sighted, near-sighted, best-first search planning strategies and adjusting the amount of planning to its cost and benefits. Moreover, some people seemed to learn from self-generated pseudo-rewards for the value of information, and some people seemed to deliberate about the value of termination whereas others do not. This suggests the prevalence of inter-individual differences, which can be subject for future work. In addition, the high proportion of non-learning models despite filtering for learning participants indicates that there is still room for improvement. For example, planning might incur cognitive costs above and beyond the cost of acquiring information (Felso, Jain, & Lieder, 2020; Callaway et al., 2022). Therefore, further work can improve our models by incorporating these additional costs into the reward signals that the models learn from.

References

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning* (pp. 115–123).

Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: Self-control and meta-decision making. *Trends in cognitive sciences*, *19*(11), 700–710.

Callaway, F., Gul, S., Krueger, P., Griffiths, T. L., & Lieder, F. (2018). Learning to select computations. In *Uncertainty in artificial intelligence: Proceedings of the thirty-fourth conference.*

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. (2018). A resource-rational analysis of human planning. In *Cogsci.*

Callaway, F., Lieder, F., Krueger, P., & Griffiths, T. L. (2017). Mouselab-mdp: A new paradigm for tracing how people plan.

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Lieder, F., & Griffiths, T. L. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, 1–14.

Daw, N. D. (2018). Are we of two minds? *Nature Neuroscience*, *21*(11), 1497.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, *112*(4), 912.

Felso, V., Jain, Y. R., & Lieder, F. (2020, July). Measuring the costs of planning. In *Proceedings of the 42nd annual meeting of the cognitive science society.* Cognitive Science Society.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(Supplement 3), 15647–15654.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, *29*, 24–30.

Hay, N. J. (2016). *Principles of metalevel control*. Unpublished doctoral dissertation, UC Berkeley.

He, R., Jain, Y. R., & Lieder, F. (2021a, December). Have i done enough planning or should i plan more? In *Workshop on metacognition in the age of ai. thirty-fifth conference on neural information processing systems.* Long Paper.

He, R., Jain, Y. R., & Lieder, F. (2021b). Measuring and modelling how people learn how to plan and how people adapt their planning strategies the to structure of the environment. In *International conference on cognitive modeling.* Retrieved from `https://mathpsych.org/presentation/604#/document`

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.

Jain, Y. R., Callaway, F., Griffiths, T. L., Dayan, P., He, R., Krueger, P. M., & Lieder, F. (2022). A computational process-tracing method for measuring people's planning strategies and how they change over time. *Behavior Research Methods*. doi: https://doi.org/10.3758/s13428-022-01789-5

Jain, Y. R., Callaway, F., & Lieder, F. (2019). Measuring how people learn how to plan. In *Cogsci* (pp. 1956–1962).

Jain, Y. R., Gupta, S., Rakesh, V., Dayan, P., Callaway, F., & Lieder, F. (2019, September). How do people learn how to plan?.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krueger, P. M., Lieder, F., & Griffiths, T. (2017). Enhancing metacognitive reinforcement learning using reward structures and feedback. In *Cogsci.*

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762–794.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS computational biology*, *14*(4), e1006043.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological review*, *126*(2), 292.

Morris, A. (2022). *Habits of thought: Model-free reinforcement learning over cognitive operations*. Unpublished doctoral dissertation.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Papernot, N., Thakurta, A., Song, S., Chien, S., & Erlingsson, Ú. (2021). Tempered sigmoid activations for deep learning with differential privacy. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 9312–9321).

Penny, W. D., Stephan, K. E., Daunizeau, J., Rosa, M. J., Friston, K. J., Schofield, T. M., & Leff, A. P. (2010). Comparing families of dynamic causal models. *PLoS computational biology*, *6*(3), e1000709.

Rieskamp, J., & Otto, P. E. (2006). Ssl: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage*, *84*, 971–985.

Rushworth, M. F., Kolling, N., Sallet, J., & Mars, R. B. (2012). Valuation and decision-making in frontal cortex:

one or many serial or parallel systems? *Current opinion in neurobiology*, *22*(6), 946–955.

Russell, S., & Wefald, E. (1991). Principles of metareasoning. *Artificial intelligence*, *49*(1-3), 361–395.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3-4), 279–292.

Wehrens, R., Putter, H., & Buydens, L. M. (2000). The bootstrap: a tutorial. *Chemometrics and intelligent laboratory systems*, *54*(1), 35–52.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3), 229–256.