How do people learn how to plan?

Yash Raj Jain, Sanit Gupta, Vasundhara Rakesh
Rationality Enhancement Group, MPI for Intelligent Systems
Tübingen, BW, Germany

Peter Dayan
MPI for Biological Cybernetics
Tübingen, BW, Germany

Frederick Callaway
Dept. of Psychology, Princeton University
Princeton, NJ, USA

Falk Lieder (falk.lieder@tuebingen.mpg.de)
Rationality Enhancement Group, MPI for Intelligent Systems
Tübingen, BW, Germany

Abstract

How does the brain learn how to plan? We reverse-engineer people's underlying learning mechanisms by combining rational process models of cognitive plasticity with recently developed empirical methods that allow us to trace the temporal evolution of people's planning strategies. We find that our Learned Value of Computation model (LVOC) accurately captures people's average learning curve. However, there were also substantial individual differences in metacognitive learning that are best understood in terms of multiple different learning mechanisms – including strategy selection learning. Furthermore, we observed that LVOC could not fully capture people's ability to adaptively decide when to stop planning. We successfully extended the LVOC model to address these discrepancies. Our models broadly capture people's ability to improve their decision mechanisms and represent a significant step towards reverse-engineering how the brain learns increasingly effective cognitive strategies through its interaction with the environment.

Keywords: decision-making; reinforcement learning; cognitive plasticity; metacognitive reinforcement learning

Introduction

One of the most distinctive aspects of human intelligence is the brain's ability to learn how to think. A better understanding of the mechanisms underlying metacognitive learning would be an important step towards building general artificial intelligence and designing interventions for improving the human mind. Here, we investigate how people learn how to plan. In previous work, we proposed two alternative models of how people learn how to think. According to the rational metareasoning model of strategy selection learning (RSSL), people learn to predict the performance of alternative decision strategies from features of the situation. By contrast, according to the Learned Value of Computation (LVOC) model, people discover and continuously change their strategy for a given environment by learning to predict the value of alternative planning operations. According to yet another proposal, the brain tweaks its metacognitive policy directly by following its performance gradient. We instantiate these general principles into concrete models of how people learn how to plan across a series of different sequential decision problems with some common characteristics. We then test these models against fine-grained process-tracing data of how people's planning strategies evolve over time.

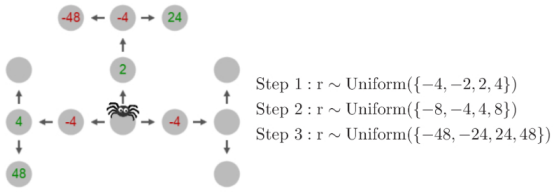

Figure 1: The Mouselab-MDP paradigm.

Methods

The Mouselab-MDP paradigm

Planning, like all cognitive processes, cannot be observed directly and must be inferred from observable behavior. This is generally an ill-posed problem. To address this challenge, we have developed a *process-tracing* paradigm that connects people's planning processes to an observable behavioral signature. In the Mouselab-MDP paradigm (Callaway, Lieder, Krueger, & Griffiths, 2017; Callaway et al., 2018), illustrated in Figure 1, participants are presented with a series of route planning problems in which each location (the gray circles) harbors a gain or loss. On each trial, participants choose among the six possible paths to maximize the total reward they receive across the three locations on the path. Initially, these rewards are occluded; however, participants can reveal the reward at a each location by clicking on it. This explicit clicking action corresponds to evaluating the quality of a future state, a fundamental cognitive operation in planning. The cognitive cost of this operation is externalized by an explicit cost of one point for each reward revealed. Here, we use the data collected by (Lieder, 2018) in which the reward structure favors a counter-intuitive backward planning strategy (Figure 1, right).

Models of metacognitive reinforcement learning

To test the previously proposed abstract principles of metacognitive reinforcement learning, we instantiate them into concrete computational models that predict the clicks participants make in Mouselab-MDP. As a first step, we have to specify how each model represents the space of possible decision-mechanisms over which learning operates.

Representation of planning strategies Jain, Callaway, and Lieder (2019) found that the participants in our experiment (Lieder, 2018) used 38 distinct planning strategies. The five most common strategies were acting without any planning (*No planning*), acting right after inspecting one immediate outcome (*Myopic Impulsive*), inspecting only the immediate outcomes and choosing the first path that starts out positively (*Myopic Satisficing*), inspecting final outcomes and acting as soon as a positive one is uncovered (*One final outcome*), and the optimal goal setting strategy that searches for and goes after the best final outcome if there is one and otherwise seeks to distinguish the the paths with the highest uncovered final outcomes (*Goal Setting*). Each of the strategies people use in the Mouselab-MDP paradigm can be expressed in terms of a weighted combination of features. And each participant's learning trajectory can be described in terms of how the weights of those features evolve over time. Some of the features that were most important to represent the participants' learning trajectories include habitual features, such as the number of times a particular branch, node, or level had been clicked on before, as well as Pavlovian features, such as whether the node lies on one of the most promising paths, as well as model-based features, such as estimate of the uncertainty about the node's value, as well as features that capture satisficing, e.g. by assigning an increasingly lower value to continuing as better paths are identified (soft-satisficing), features that govern pruning (e.g., a feature that assigns a negative value to thinking about paths whose expected value is -24 or less, and other features. In the models reported below all features were normalized to lie between 0 and 1. For a detailed documentation of all 56 features and all 38 strategies please see `https://osf.io/hakbz/`.

Learning mechanisms Having specified the models' representation of decision strategies, we now specify three models of the learning mechanism that might operate on these representations: rational strategy selection learning (Lieder & Griffiths, 2017, RSSL), learning the value of computation (Krueger, Lieder, & Griffiths, 2017, LVOC), and strategy discovery by gradient ascent over the space of decision mechanisms according to the classic REINFORCE algorithm (Williams, 1992, REINFORCE). In order to compare the learning that happens in these models to learning in people, we i) fix the click sequence on the first trial to the clicks performed by the participant, ii) fit each model's prior on feature weights or strategies to each participant's data individually, and iii) simulate each participant's click sequences by applying the fitted model to the exact sequence of 31 planning problems the participant was given.

The RSSL model treats the problem of deciding how to plan as a 38-armed bandit with one arm for each strategy. It performs Bayesian inference on the expected return of each strategy and selects strategies via Thompson sampling. It has $38 \times 2 = 76$ free parameters that specify the prior mean and variance of each strategy's expected return.

The LVOC model learns an approximation $\hat{Q}_{\text{meta}} = \sum_{i=1}^{56} w_i \cdot f_i(b,c)$ to the value $Q_{\text{meta}}(b,c)$ of performing planning operation c in belief state b via a Bayesian version of the SARSA temporal difference learning algorithm (Krueger et al., 2017). The planning operations correspond to clicking or terminating planning, and the belief state is determined by which payoffs have already been observed and what their values are. The LVOC model has two free parameters: the variance σ^2_{prior} of its prior $\mathcal{N}(\mathbf{w}; \mu_{\text{prior}}, \sigma^2 \cdot \text{Id})$ on the weights $\mathbf{w}$ and the number of samples it draws from its posterior on $\mathbf{w}$ to predict Q_{meta}. To capture individual differences, we fit these parameters to process-tracing data from individual participants. The mean vector μ_{prior} was initialized with the feature weights of the strategy participants used on the first trial according to our computational microscope (Jain et al., 2019).

The REINFORCE model applies the vanilla version of the policy gradient algorithm REINFORCE to learn the parameters θ of a softmax policy $\pi_\theta(c|b) \propto \exp\left(\frac{1}{\tau} \cdot \sum_{k=1}^{56} \theta_k \cdot f_k(b,c)\right)$ for selecting planning operations based on the features $\mathbf{f}$ described above. Its learning rate is optimized online using ADAM and its discount factor γ and decision temperature τ are free parameters.

Model selection

We use leave-one-out cross-validation (Friedman, Hastie, & Tibshirani, 2001) to estimate the generalization error of each model's predictions of i) the learning curves of individual participants, ii) individual participants' time series of the feature weights, and iii) the time series of individual participants' strategies. The participants' time series of weights and strategies were determined by inverting a generative model of how people's strategies manifest in their click sequences (Jain et al., 2019). Leave-one-out cross-validation was performed by training the model on the data from all but one trial and predicting the criterion variable on the left-out trial. Each trial was left out once and then a distance metric between the predicted and the observed value on the left-out trials were averaged across all folds. Model selection was performed separately for each participant by selecting the model that had the lowest generalization error in predicting the criterion variable on average across the 31 cross-validation folds.

Results

Figure 2 shows the average learning curves of people versus the fitted models. The LVOC model captured the average of peoples learning curves remarkably well. By contrast, the REINFORCE model was unable to capture the full extent of people's metacognitive reinforcement learning. On average, the RSSL model learned faster than people and achieved a higher average performance than the average participant.

Clustering participants based on their initial performance (avg. score on trials 1–5) and their final performance (avg. score on trials 22–31) revealed substantial individual differences. Model selection based on the AIC suggested that there were three types of participants: i) 9 participants started

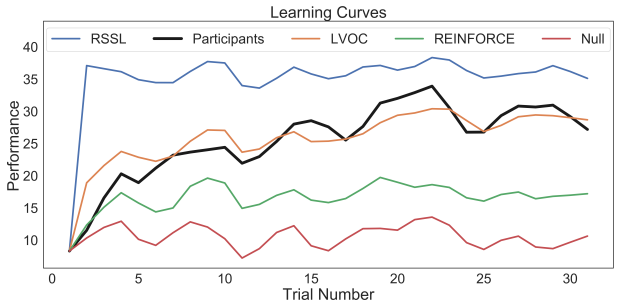

Figure 2: Comparison of learning curves. Each learning curve shows the participants' average score (sum of payoffs minus planning cost) in the 1st, 2nd, 3rd, $\cdots$, 31st Mouselab-MDP planning problem (see Figure 1)

.

out performing poorly and did not improve (-6.1 points/trial vs. -1.5 points/trial), ii) 5 participants started out performing poorly and improved significantly (-14.7 points/trial vs. 39.8 points/trial), and iii) 25 participants performed well from the beginning (28.8 points/trial vs. 39.2 points/trial).

Model Comparison Results

We evaluated each of the proposed models against a baseline model that just repeats the participant's initial strategy. (m_{null}). Given the substantial individual differences, we performed model selection separately for each group of participants (in addition to performing model selection across all participants). Table 1 summarizes the results of within-subject model selection by criterion. These results show that the individual differences in people's performance can be understood in terms of different learning mechanisms: The data from the consistently low-scoring participants (Table 1a) was best explained by the null model (no learning) whereas the data from participants who improved substantially (Table 1b) was best explained by the LVOC model, and the data from the constantly high-performing participants (Table 1c) was best explained by the RSSL model. These findings suggest that a substantial source of individual differences in performance is that some people enter the task with effective planning strategies whereas others do not engage in learning at all. Among participants who did improve significantly (Group 2), learning was best captured by the strategy discovery mechanisms of the LVOC model.

Across all participants, the LVOC model was best at predicting the temporal evolution of the weights that people's decision strategies assigned to the different features, and the RSSL model was best at predicting their strategy sequences and their performance. Given that some aspects of people's learning were best accounted by learning when to select which strategy whereas others were best accounted by tuning the weights that define the strategy, it is conceivable that strategy selection learning and strategy discovery learning jointly shape how people learn how to plan.

a) Consistently low-scoring participants ($n = 9$)

	m_{null}	REINFORCE	RSSL	LVOC
performance	**5**	1	1	2
strategies	2	**3**	**3**	1
weights	3	0	0	**6**

b) Participants who improved substantially ($n = 5$)

	m_{null}	REINFORCE	RSSL	LVOC
performance	0	1	1	**3**
strategies	1	0.5	0	**3.5**
weights	1	0	**2**	**2**

c) Consistently high-scoring participants ($n = 25$)

	m_{null}	REINFORCE	RSSL	LVOC
performance	5.33	2.33	**13.33**	4
strategies	8	0.5	**14**	2.5
weights	5.5	5.5	**7**	**7**

d) All participants ($n = 39$)

	m_{null}	REINFORCE	RSSL	LVOC
performance	10.33	4.33	**15.33**	9
strategies	11	4	**17**	7
weights	9.5	5.5	9	**15**

Table 1: Number of participants whose data was best explained by each of the models broken down by model selection criterion and group.

Shortcomings of the LVOC model

The model comparisons reported above often favored the RSSL model, which assumes that participants make discrete shifts between pre-defined planning strategies. However, this model cannot capture the gradual learning curve that many participants demonstrated. These participants who started from a low-performing strategy took about 7.5 ± 1.4 trials to transition to a high-performing strategy whereas the RSSL model accomplishes this transition in only 2.6 ± 0.1 trials on average. The LVOC model captures this gradual increase, but often lost to the RSSL model in the model comparisons at the level of individual participants. To understand why, we investigate which aspects of how people's planning strategies evolve over time the LVOC model might be failing to capture.

We found that the LVOC model captured the gradual decrease of acting impulsively and myopic satisficing, and also the gradual increase of the optimal goal-setting strategy. But its predictions deviated from people's performance in the following ways: First, while the prevalence of the myopic impulsive strategy gradually decreased among people, the LVOC model predicted that its frequency should initially increase and then return to its original level. Second, even when fit to individual participants' strategy sequences, the LVOC succeeded less frequently in improving upon its initial strategy than people (69% vs. 97%). Third, LVOC discovered the optimal goal setting strategy less often than people (11.8% vs. 54%). Instead, the LVOC model most often learned to use strategies that combine goal-setting with unnecessary extraneous planning (i.e., the *Best Final Outcome* strategy and the *Immediate Outcomes and Goal Setting* strategy). This suggests that while the LVOC model can learn to prioritize long-term consequences, it does not fully capture people's ability to learn an adaptive stopping rule.

Modeling how people decide when to stop planning

The primary shortcoming of the basic LVOC model evaluated above is that it continues planning long after people would stop planning. To address this shortcoming, we investigated two potential sources of this discrepancy: i) planning might incur an additional computational cost for people that the basic LVOC model does not capture, and ii) people might employ a simple stopping rule to decide when to terminate planning. To investigate the first possibility, we developed an extended version of the LVOC model that includes an additional cost parameter that is subtracted from its meta-level reward for executing a planning operation. To test the second possibility we developed two additional models.

Two-Stage Models Previous findings indicate that foraging decisions involve two separate decision systems: while the ventromedial prefrontal cortex appears to estimate and compare the values of alternative options the dorsal anterior cingulate cortex appears to decide whether to continue collecting more information (e.g., by foraging) or to choose the best option found so far (Rushworth, Kolling, Sallet, & Mars, 2012). Since deciding how to plan is like foraging for information, it might be implemented via two separate systems as well. We therefore develop a model that first decides whether to continue planning (Stage 1) and then either selects the action that looks best so far or selects the next planning operation according to the LVOC model (Stage 2). We instantiated this idea in two 2-stage models that use the stopping rules that (according to the AIC) best explained the stopping behavior of the largest and the second largest number of participants respectively. Concretely, the probabilistic stopping rule that was the best model for the largest number of participants (i.e., 18/39) was $P(C_t = \text{stop}|b_t) = \text{sigmoid}\left(1/\tau \cdot \left(\max_{\text{path}} \mathbb{E}\left[R(\text{path})|b_t\right] - \theta\right)\right)$, where the free parameters θ and τ correspond to the decision-maker's aspiration level and noisiness respectively. The decision rule that was best for the second largest number of participants (i.e., $10/39$) was an extension of this rule where the aspiration level θ gradually decreases with the number of clicks the participant has already made (i.e., $\theta = a - b \cdot n_{\text{clicks}}$).

Results The addition of a second decision stage reduced the LVOC model's generalization error in predicting people's performance, allowed it to capture that Optimal Goal Setting becomes the most prevalent planning strategy (with a prevalence of 31.9% and 29.4% respectively), and increased its propensity to improve on the initial strategy to a near-human level (86.9% and 84.6% respectively). The addition of a cost parameter neither increased nor decreased the LVOC model's generalization error. Most importantly, according to the AIC, the three new models provided the best explanation for the performance of 17 out of 39 participants. Taken together, the four LVOC models provide the best explanation for the majority of all participants' learning curves ($20/39$). The data from the remaining participants were best explained by the RSSL model (12.5/39), the null model (4.5/39), and the REINFORCE model (2/39). The proportion of participants whose weights were best explained by one of the LVOC models was even higher (23/39). Out of the four LVOC models, the 2-stage model with an adaptive stopping criterion performed best at predicting people's performance (8/39) and the basic LVOC model performed best at predicting the temporal evolution of the feature weights (9/39). Furthermore, the LVOC models best explained the scores of low-performing participants and substantially-improving participants and the feature-weights of consistently high-performing participants, whereas the RSSL model remained the best explanation for the performance of the consistently high-performing participants.

Additional results can be found at `https://osf.io/hakbz/`.

Conclusion

Given that this is the first time that metacognitive reinforcement learning has been modelled at the level of individual computations, the LVOC model got surprisingly close to the average of people's learning curves. Furthermore, the LVOC model and its extensions jointly provided the best explanations for the majority of our participants' data. Despite the success of the LVOC model, we also observed considerable individual differences in planning that might partly result from different people relying on different learning mechanisms and different rules for deciding when to stop planning. While the strategy discovery mechanism of the LVOC model is important, additional mechanisms – such as selecting strategies from a toolbox of pre-existing strategies – might also be necessary to capture human learning. Future work will therefore investigate how strategy selection and strategy discovery interact in human metacognitive learning, evaluate improved versions of the RSSL model, and further elucidate individual differences in how people learn how to plan.

References

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

Callaway, F., Lieder, F., Krueger, P. M., & Griffiths, T. L. (2017). Mouselab-mdp: A new paradigm for tracing how people plan. In *The 3rd multidisciplinary conference on reinforcement learning and decision making*.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York.

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

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

Lieder, F. (2018). Developing an intelligent system that teaches people optimal cognitive strategies. In F. Lieder (Ed.), *Beyond bounded rationality: Reverse-engineering and enhancing human intelligence* (chap. 8).

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

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.

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