# Discovering and Teaching Optimal Planning Strategies

**Falk Lieder[1,2,*], Frederick Callaway[1,*], Paul M. Krueger[1], Priyam Das[1], Thomas L. Griffiths[1], Sayan Gul[1]**
[1] UC Berkeley, [2] MPI for Intelligent Systems, Tübingen [*] Equal contribution.

## Abstract

How should we think and decide, and how can we learn to make better decisions? To address these questions we formalize the discovery of cognitive strategies as a metacognitive reinforcement learning problem. This formulation leads to a computational method for deriving optimal cognitive strategies and a feedback mechanism for accelerating the process by which people learn how to make better decisions. As a proof of concept, we apply our approach to develop an intelligent system that teaches people optimal planning stratgies. Our training program combines a novel process-tracing paradigm that makes peoples latent planning strategies observable with an intelligent system that gives people feedback on how their planning strategy could be improved. The pedagogy of our intelligent tutor is based on the theory that people discover their cognitive strategies through metacognitive reinforcement learning. Concretely, the tutor's feedback is designed to maximally accelerate people's metacognitive reinforcement learning towards the optimal cognitive strategy. A series of four experiments confirmed that training with the cognitive tutor significantly improved people's decision-making competency: Experiment 1 demonstrated that the cognitive tutor's feedback accelerates participants' metacognitive learning. Experiment 2 found that this training effect transfers to more difficult planning problems in more complex environments. Experiment 3 found that these transfer effects are retained for at least 24 hours after the training. Finally, Experiment 4 found that practicing with the cognitive tutor conveys additional benefits above and beyond verbal description of the optimal planning strategy. The results suggest that promoting metacognitive reinforcement learning with optimal feedback is a promising approach to improving the human mind.

**Introduction** Research on heuristics and biases has identified many ways in which human judgment and decision-making might be sub-optimal (1). This sub-optimality presents an opportunity for improvement. Two of the main challenges for improving the human mind are a) discovering effective cognitive strategies, and b) teaching them in a way that people will apply them in everyday life. Here, we address both challenges by leveraging two recent advances in computational cognitive science. To address the first problem, we derive resource-rational strategies (2; 3). To address the second problem, we build on the idea that people learn to use near optimal cognitive strategies through metacognitive reinforcement learning (4; 5). We combine these ideas to develop a feedback mechanism for teaching people optimal cognitive strategies.

**A cognitive tutor for teaching people optimal planning strategies** As a proof-of-concept, we applied our approach to teaching people optimal planning strategies in the Mouselab-MDP paradigm (6), illustrated in Figure 1a. On each trial, participants solve a route planning problem where each location (the gray circles in Figure 1) harbors a reward or punishment. The participant must travel to one of the outer locations, ideally maximizing the rewards attained on the way. Critically, these potential gains and losses are initially occluded; the participant must click on a location to reveal its value. In this way, the unobservable cognitive operations involved in planning (e.g. mental simulation and recollection) are externalized as observable behavior. To capture the resource costs of these operations, each click incurs a fixed fee.

To discover optimal planning strategies we formalized planning as a meta-level MDP (7) and solved the meta-level MDP using backward induction (8). This yields the optimal planning strategy in the form of a meta-level value function $Q_{\text{meta}}$ which encodes the expected long term value of clicking on a location given the currently available information. The cognitive tutor uses this value function to provide pseudo-rewards in the form of delays, following the reward shaping method (9) that was originally developed to accelerate model-free reinforcement learning in robots. Additionally, the tutor demonstrates what the optimal planning strategy, would have done instead (Figure 1b). We applied this approach to multi-step decision problems where the variance of the rewards increases from each step to the next. For this environment, the optimal strategy is to first set a goal by evaluating potential end states.

**Evaluation** We evaluated the cognitive tutor in four experiments. Experiments 1-3 employed a between-subjects pre-post design comparing the effects of practicing the task with versus without the cognitive tutor. In Experiment 1, the pre-test, training, and post-test blocks all employed the same 3-step planning task. We found that the tuitor's metacognitive feedback significantly accelerated our participant's learning (see Figure 1c) and led to significantly higher post-test performance ($36.2$ \$/trial vs. $24.6$ \$/trial, $t(2258) = 10.7$, $p < 0.0001$). In Experiment 2, the training block used a flight planning task that was structurally equivalent to task used in Experiment 1, but the pre-test and post-test blocks used the transfer task shown in Figure 2a). As shown in Figure 2b), we found significant transfer effects from the relatively simple 3-step training task to the more difficult 5-step transfer task. This transfer effect was mediated by people learning to plan backward – which was also beneficial in the transfer task. Experiment 3 was like Experiment 2 except that there was an approximately 24 h delay

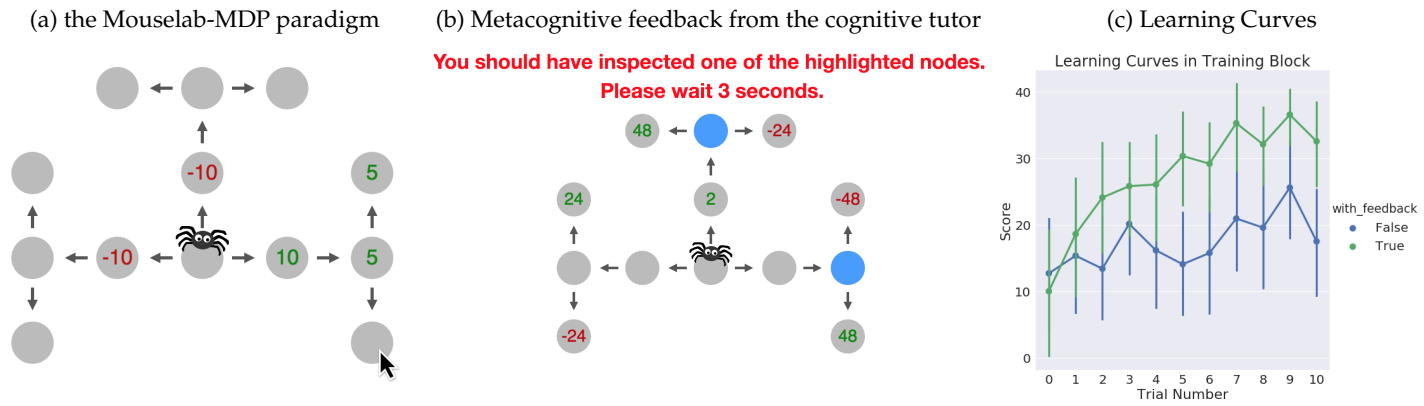


Figure 1: The cognitive tutor accelerates learning.

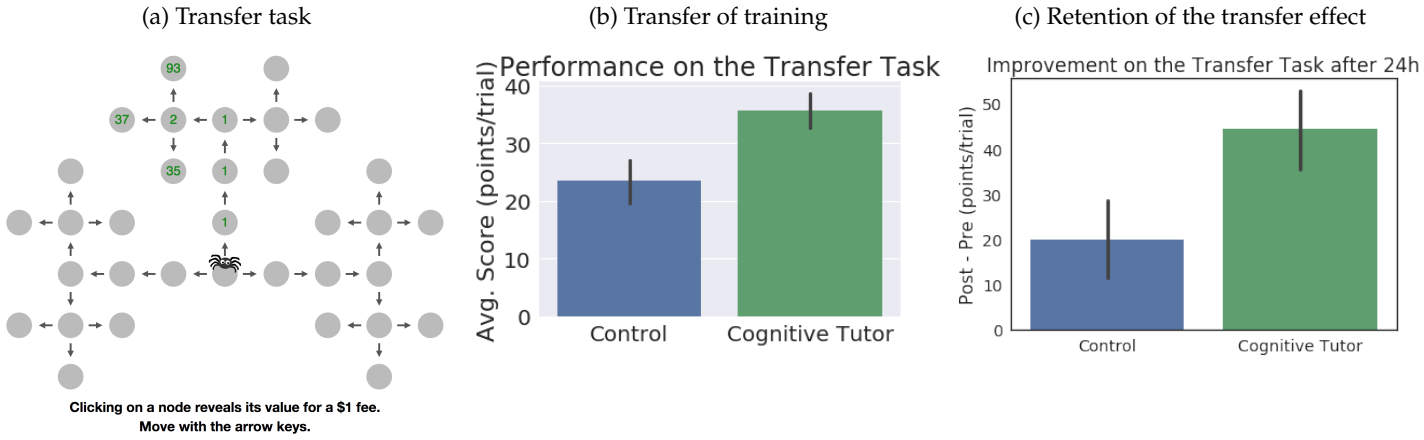


Figure 2: Transfer experiments.

between the training block and the post-test block. As shown in Figure 2c), we found that the transfer effect was retained. Experiment 4 compared the effectiveness of instruction plus practice with the cognitive tutor versus pure instruction and instruction plus demonstration. Participants in all three conditions read about the goal-setting principle for better decision-making discovered by our method. In the experimental conditions participants subsequently practiced applying the goal-setting principle with the cognitive tutor or saw a video demonstration of the optimal strategy. After 24 h all participants were tested on the same transfer task. We found that participants who had practiced with the cognitive tutor performed significantly better on the transfer task than participants who were only told about the principle ($38.0\,\$/\text{trial}$ vs. $24.2\,\$/\text{trial}$, $t(83) = 10.5, p = 0.0000$). Participants who had seen a demonstration of the optimal strategy performed at the same level as participants who had practiced with the cognitive tutor ($38.8\,\$/\text{trial}$ vs. $38.0\,\$/\text{trial}$, $t(78) = -0.7, p = 0.49$).

**Discussion** The theoretical framework of resource-rationality allowed us to derive near-optimal planning strategies automatically, and the theory of metacognitive reinforcement learning allowed us to develop an intelligent system that can teach those rational heuristics very effectively. Our preliminary results suggest that practice with our cognitive tutor is more effective than instruction and has transferable benefits that are retained over time. This suggests that promoting metacognitive reinforcement learning with optimal feedback is a promising approach to enhancing human rationality.

## References


[1] A. Tversky, D. Kahneman, *Science* **185**, 1124 (1974).

[2] F. Lieder, P. M. Krueger, T. L. Griffiths, *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, G. Gunzelmann, A. Howes, T. Tenbrink, E. J. Davelaar, eds. (Cognitive Science Society, Austin, TX, 2017), pp. 742–747.

[3] F. Lieder, F. Callaway, S. Gul, P. M. Krueger, T. L. Griffiths, *NIPS workshop on Cognitively Informed AI* **abs/1711.06892** (2017).

[4] P. M. Krueger, F. Lieder, T. L. Griffiths, *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (Cognitive Science Society, 2017).

[5] F. Lieder, T. Griffiths, *Psychological Review* **124**, 762 (2017).

[6] F. Callaway, F. Lieder, P. M. Krueger, T. L. Griffiths, *The 3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making, Ann Arbor, MI* (2017).

[7] N. Hay, S. Russell, D. Tolpin, S. Shimony, *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, N. de Freitas, K. Murphy, eds. (AUAI Press, Corvallis, OR, 2012).

[8] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming* (John Wiley & Sons, Hoboken, NJ, 2014).

[9] A. Y. Ng, D. Harada, S. Russell, *Proceedings of the 16th Annual International Conference on Machine Learning*, I. Bratko, S. Dzeroski, eds. (Morgan Kaufmann, San Francisco, CA, 1999), pp. 278–287.