A reward shaping method for promoting metacognitive learning

Falk Lieder[1]
Dept. of Psychology, UC Berkeley
falk.lieder@berkeley.edu

Paul M. Krueger[1]
Dept. of Psychology, UC Berkeley
pmk@berkeley.edu

Frederick Callaway[1]
Dept. of Psychology, UC Berkeley
fredcallaway@berkeley.edu

Thomas L. Griffiths
Dept. of Psychology, UC Berkeley
tom_griffiths@berkeley.edu

[1] **These authors contributed equally.**

Abstract

The human mind has an impressive ability to improve itself based on experience, but this potential for cognitive growth is rarely fully realized. Cognitive training programs seek to tap into this unrealized potential but their theoretical foundation is incomplete and the scientific findings on their effectiveness are mixed. Recent work suggests that mechanisms by which people learn to think and decide better can be understood in terms of metacognitive reinforcement learning. This perspective allow us to translate the theory of reward shaping developed in machine learning into a computational method for designing feedback structures for effective cognitive training. Concretely, our method applies the shaping theorem for accelerating model-free reinforcement learning to a meta-decision problem whose actions are computations that update the decision-maker's probabilistic beliefs about the returns of alternative courses of action. As a proof of concept, we show that our method can be applied to accelerate learning to plan in an environment similar to a grid world where every location contained a reward. To measure and give feedback on people's planning process, each reward was initially occluded and had to be revealed by clicking on the corresponding location. We found that participants in the feedback condition learned faster to deliberate more and consequently reaped higher rewards and identified the optimal sequence of moves more frequently. These findings inspire optimism that meta-level reward shaping might provide a principled theoretical foundation for cognitive training and enable more effective interventions for improving the human mind by giving feedback that is optimized for promoting metacognitive reinforcement learning.

Keywords: metacognitive reinforcement learning; meta-decision making; cognitive training; reward shaping; feedback

Acknowledgements

This work was supported by grant number ONR MURI N00014-13-1-0341 and a grant from the Templeton World Charity Foundation. The authors thank Thomas Hills, Peter Dayan, Rika Antonova, Silvia Bunge, Stuart Russell, Smitha Milli, Amitai Shenhav, and Sebastian Musslick for feedback and discussions.

1 Introduction and Background

One of the most remarkable aspects of the human mind is its ability to improve itself based on experience. Such learning occurs in a range of domains, from simple stimulus-response mappings, motor skills, and perceptual abilities, to problem solving, cognitive control, and learning itself (Green & Bavelier, 2008; Bavelier, Green, Pouget, & Schrater, 2012). Demonstrations of cognitive and brain plasticity have inspired cognitive training programs. The success of cognitive training has been mixed and the underlying learning mechanisms are not well understood (Owen et al., 2010; Anguera et al., 2013; Morrison & Chein, 2011). Feedback is an important component of many effective cognitive training programs, but it remains unclear what makes some feedback structures more effective than others, and there is no principled method for designing optimal feedback structures.

Recent work suggests that cognitive growth can be fruitfully conceptualized in terms of a metacognitive reinforcement learning mechanism that identifies which cognitive operation should be executed depending on the current mental state (Krueger, Lieder, & Griffiths, 2017). One instance of metacognitive reinforcement learning is learning how to decide. This includes learning when to rely on habits versus planning and learning how far to plan ahead. Concretely, Krueger et al. (2017) found that an approximate Q-learning algorithm qualitatively captured how quickly people learn to plan farther ahead depending on the reward structure of their training problems. This perspective allows us to translate methods for accelerating reinforcement learning in robots (Ng, Harada, & Russell, 1999) into feedback structures for cognitive training in humans.

Reinforcement learning problems are commonly modeled as Markov decision processes (MDPs) which are defined by a set of states S, a set of actions A, transition probabilities T, a reward function r, and a discount factor γ (Sutton & Barto, 1998). The goal is to learn a policy $\pi : S \mapsto A$ that attains the highest possible sum of (discounted) rewards $V^{\star}(s)$ that can be attained starting from any state s. Ng et al. (1999) showed the model-free reinforcement learning can be accelerated without changing the optimal policy by adding pseudo-rewards (PR) of the form

$$\text{PR}(s, a, s') = \gamma \cdot \hat{V}(s') - \hat{V}(s) \tag{1}$$

to the reward function r where $\hat{V}$ is an approximation to $V^{\star}$. For instance, a robot learning to navigate a maze may receive a positive PR for moving towards its target location and a negative PR for moving away from it. The resulting reward structure $r'(s, a, s) = r(s, a, s') + \text{PR}(s, a, s')$ makes the optimal policy easier to learn because it aligns immediate reward with long-term value.

Previous work suggested that these pseudo-rewards can also be used to improve human decision-making (Lieder & Griffiths, 2016). Concretely, Lieder and Griffiths (2016) found that presenting pseudo-rewards as incentives increases decision-quality. While this intervention aligned people's choices with the optimal policy, it is unlikely to have improved their planning process in ways that would benefit them in novel environments, because the incentives eliminated the need for planning. Hence, to promote learning to plan it might be more effective to let people plan how to solve the original problem and provide feedback on their planning process. Providing such feedback is challenging for two reasons: first, planning processes cannot be observed directly, and second planning performance is difficult to score automatically. To overcome this first challenge, we leverage the newly developed Mouselab-MDP paradigm (Callaway, Lieder, Krueger, & Griffiths, 2017) that makes it possible to measure people's planning processes by tracing the information they acquire during planning (see Figure 1). To overcome the second challenge, we model optimal planning as the solution to a meta-level MDP (Hay, Russell, Tolpin, & Shimony, 2012) whose actions are cognitive operations and whose states are the planner's beliefs about the rewards it can harvest from the environment by taking a certain action in a certain state. A meta-level MDP M formalizes the problem of selecting the optimal sequences of computations for a resource-bounded computational architecture. This is known as rational metareasoning in the artificial intelligence literature (Russell & Wefald, 1991). Concretely, a meta-level MDP

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

is a Markov decision process whose actions $\mathcal{C}$ are cognitive operations, states $\mathcal{B}$ represent the agent's probabilistic beliefs, and transition function T_{meta} models how cognitive operations change the agent's beliefs. In addition to a set of computations C that update the agent's belief, the cognitive operations also include the meta-level action $\perp$ that terminates deliberation and translates the current belief into action. The meta-level state b_t encodes the agent's probabilistic beliefs about the domain it is reasoning about. The meta-level reward function r_{meta} captures the cost of thinking and the reward r the agent expects to receive from the environment when it stops deliberating and takes action. The computations C do not yield any external reward. Their only effect is to update the agent's beliefs. Hence, the meta-level reward for performing a computation $c \in C$ is $r_{\text{meta}}(b_t, c) = -\text{cost}(c)$. By contrast, terminating deliberation and taking action ($\perp$) does not update the agent's belief. Instead, its value lies in the anticipated reward for taking action, that is $r_{\text{meta}}(b_t, \perp) = \arg\max_a b_t^{(\mu)}(a)$, where $b_t^{(\mu)}(a)$ is the expected reward of taking action a according to the belief b_t.

Here, we apply reward-shaping to a meta-level MDP of the planning process to derive an optimal feedback structure for accelerating metacogntive reinforcement learning in people.

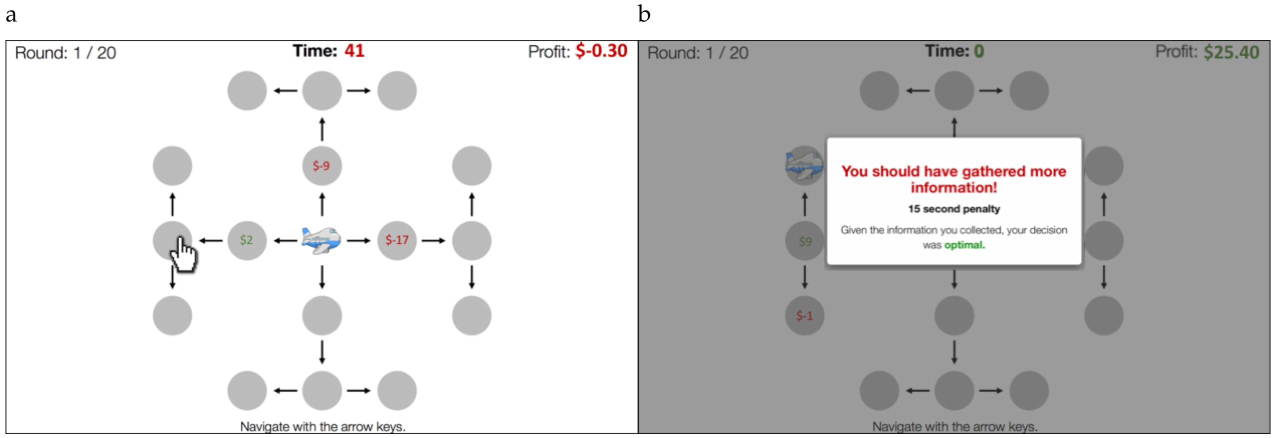

Figure 1: Illustration of Mouselab-MDP paradigm used to trace and give feedback on how people plan. Participants plan a sequence of three moves so as to maximize the sum of rewards along the path while minimizing the cost of planning. Each trial starts with a new set of occluded rewards. To observe these rewards participants have to click on them (a) and are provided with feedback after each flight (b). Each click costs 10 cents.

2 A Method for Designing Optimal Feedback Structures for Metacognitive RL

Building on the research summarized above, our method for designing optimal feedback structures to promote metacognitive learning proceeds in the following four steps:

1. Model the cognitive function to be improved (e.g., planning) and the available cognitive operations (e.g., determining the reward for taking a certain action in a certain state) and their costs as a meta-level MDP.
2. Approximate the optimal value function $V^\star_{\text{meta}}$ of the meta-level MDP.
3. Translate the approximate meta-level value function $\hat{V}_{\text{meta}}$ into integrated meta-level pseudo-rewards

$$\text{PR}(b_t, c_t) = \mathbb{E}\left[\hat{V}_{\text{meta}}(B_{t+1})|b_t\right] - \hat{V}_{\text{meta}}(b_t) + \mathbb{E}\left[r_{\text{meta}}(b_t, c, b_{t+1})\right]$$

4. Translate these PRs into rewards and punishments that provide meaningful feedback to people (e.g., Figure 1b).

This method is very general and widely applicable. Here we apply it to accelerating the process by which people learn to plan in the Mouselab-MDP paradigm (Callaway et al., 2017). In this paradigm the task is to find a sequence of moves (m) to collect as much reward as possible from the locations (l) along the resulting path (see Figure 1a). Critically, all of the rewards r are initially occluded. They can be uncovered by clicking on their location and each click has a cost.

First, we model deciding how to plan under cognitive constraints as a meta-level MDP. The set of cognitive operations $\mathcal{C}$ comprises the computations $C = \{c_1, \cdots, c_n\}$, each of which retrieves the reward of one of the n locations (e.g., by clicking on one of the locations in Figure 1a) and updates the belief state accordingly, and the operation $\perp$ which executes the move with the highest expected return according to the current belief state b_t. The belief state $b_t \in \mathcal{B}$ is defined by the means $\mu_{l,m}$ and variances $\sigma_{l,m}$ of normal distributions on the Q-values $Q(l, m)$ of performing move m in location l of the environment shown in Figure 1 and the current location l. The probability of the next belief state b_{t+1} given the current belief state b_t and a computation c_t, that is $T_{\text{meta}}(b_t, c_t, b_{t+1})$, is defined by the assumptions that the reward of the inspected location will be sampled from the reward distribution $P(R) = \mathcal{N}(\mu_R, \sigma_R)$ and the belief state will be updated according the the laws of probability theory. The meta-level reward $r_{\text{meta}}(b_t, c_t)$ of performing a computation was set to a constant $-\lambda$ which represents the cost of generating and processing another piece of of information. The meta-level reward for acting is the expected reward of the resulting action minus the cost of planning, that is $r_{\text{meta}}(b_t, \perp) = \mathbb{E}\left[r(l, \arg\max_m \mu^{(b_t)}_{l,m})\right] - \tau \cdot \sum_{\mathbf{m} \in M_l} \#\mathbf{m}$,where τ is the planning cost per step, M_l is the set of paths starting from location l, and $\#\mathbf{m}$ is the length of the path $\mathbf{m}$. Here, we set the parameters of this meta-level MDP such that the benefit of complete planning according to perfect information far outweigh its cost (i.e., $\mu_R = 4.50, \sigma_R = 10.60, \lambda = 0.10$, and $\tau = 0.01$). Because of the chosen parameters we can approximate the optimal meta-level value function under the assumption that the agent will deliberate enough to eventually select the best move. We thus approximate the value of performing computation c_t in belief state b_t by

$$\hat{V}_{\text{meta}}(b_t) = \mathbb{E}\left[\max_m Q(l_t, m)|b_t\right] - \lambda \cdot \#U_{b_t} - \tau \cdot \sum_{\mathbf{m} \in M_l} \#\mathbf{m}, \tag{3}$$

where $\#U_{b_t}$ is the number of unobserved locations ahead of the current location.

The resulting pseudo-rewards $\text{PR}(b_t, c_t, b_{t+1}) = \mathbb{E}\left[\hat{V}_{\text{meta}}(B_{t+1})|b_t\right] - \hat{V}_{\text{meta}}(v_t) + \mathbb{E}\left[r_{\text{meta}}(b_t, c, b_{t+1})\right]$ satisfy intuitive desiderata: They are negative when the agent decides prematurely, wastes time on irrelevant information, or takes an action that is suboptimal with respect to the current belief state. Since the rewards of the task environment (Figure 1) are monetary, the meta-level pseudo-rewards can be interpreted as the expected future monetary consequences of the cognitive operations the participant performed or failed to perform. We can therefore convert the pseudo-rewards into a delay-penalty by assuming that our participants value each hour of their time at about $\$10$. Critically, pseudo-rewards for thinking reflect the expected gain in decision quality regardless of whether it made the person more optimistic or more pessimistic. Next, we evaluate whether the resulting feedback can accelerate the rate at which people learn to plan better in the Mouselab-MDP paradigm shown in Figure 1.

3 Experimental Test of the Method's Efficacy

3.1 Methods

We recruited 60 participants on Amazon Mechanical Turk. The task took 29.7 minutes on average and participants were paid $1.50 plus a performance-dependent bonus equal to 5% of their earning from one randomly selected trial; the average bonus was $1.76. There were two experimental conditions: In the feedback condition, participants received a delay penalty computed by the reward shaping method described above along with feedback messages. In the control condition, the delay was constant regardless of performance and participants received no feedback. The assignment of participants to conditions was counterbalanced using psiTurk (Gureckis et al., 2016). All participants solved twenty planning problems that were rendered using the Mouselab-MDP plug-in for JsPsych (Callaway et al., 2017). In each problem, participants had to route an airplane from the center of the screen to one of eight final destinations, with two additional stops along the way (Figure 1a). Each state adds or subtracts money, and the value of each state is initially occluded, but can be revealed at any time by clicking on it. Each click incurs a cost of $\$0.10$. Participants were required to spend at least 45 seconds on every trial, to ensure that clicking doesn't incur a time cost.

The twenty planning problems differed only in the reward function. These reward functions were generated independently of one another and their order was randomized across participants. All rewards were integers between -20 and $+18$ with mean 4.2 and standard deviation 10.3. The reward functions were designed such that three-step planning based on complete information led to significantly higher average returns than 2-step planning, 1-step planning, and random choice ($\$38.35$ vs. $\$16.06$–$\$18.30$). One-step planning performed at chance level.

In the feedback condition, we calculated a pseudo-reward for each of the participant's clicks and moves according to meta-level reward shaping. Whenever the participant chose a move the sum of the pseudo-rewards obtained since the last move was translated into a delay penalty according to a conversion rate of $\$10$ per hour. If the resulting penalty was less than 1 second participants could move on immediately. Otherwise, the move was followed by a timeout period that lasted for the computed duration. During this period the participant could not proceed with the task, the countdown on the minimum trial duration stopped, and a feedback message was shown. As illustrated in Figure 1b, this message gave feedback on whether the participant had considered too few possible paths, whether they had considered irrelevant states, and whether their decision was suboptimal relative to what they knew at the time.

3.2 Results

We quantified how much planning our participants performed by how many of the 16 relevant locations they inspected (# clicks). The quality of the resulting decisions was measured by the relative reward $\frac{r - r_{\text{min}}}{r_{\text{max}} - r_{\text{min}}}$ (where r_{max} and r_{min} are the highest and the lowest total return achievable on that trial) and by the frequency with which participants selected the optimal sequences of moves (% optimal routes).

We found that participants in both conditions learned to collect more information and make better plans (Figure 2). To quantify these effects we fitted the learning curves $y(t)$ shown in Figure 2 by a sigmoidal function of trial number t,

$$\hat{y}(t) = (1 - \beta_0) \cdot \frac{1}{1 + \exp\left(-(\beta_1 + (\beta_2 + \beta_3 * c) * t)\right)} + \varepsilon, \tag{4}$$

where β_0 determines asymptotic performance, β_1 determines initial performance, β_2 is the learning rate in the control condition, β_3 measures how much learning is accelerated by feedback ($c = 1$ indicates the presence of feedback), and ε is an i.i.d. random perturbation. To determine whether performance improved with experience and whether this improvement was modulated by feedback we performed a nonlinear regression analysis (`fitnlm` in `Matlab 2015b`) to estimate the coefficients β and test whether β_2 and β_3 were significantly greater than zero. This analysis revealed significant effects of learning on the number of clicks ($t(1197) = 5.04, p < .0001$), the frequency of choosing the optimal

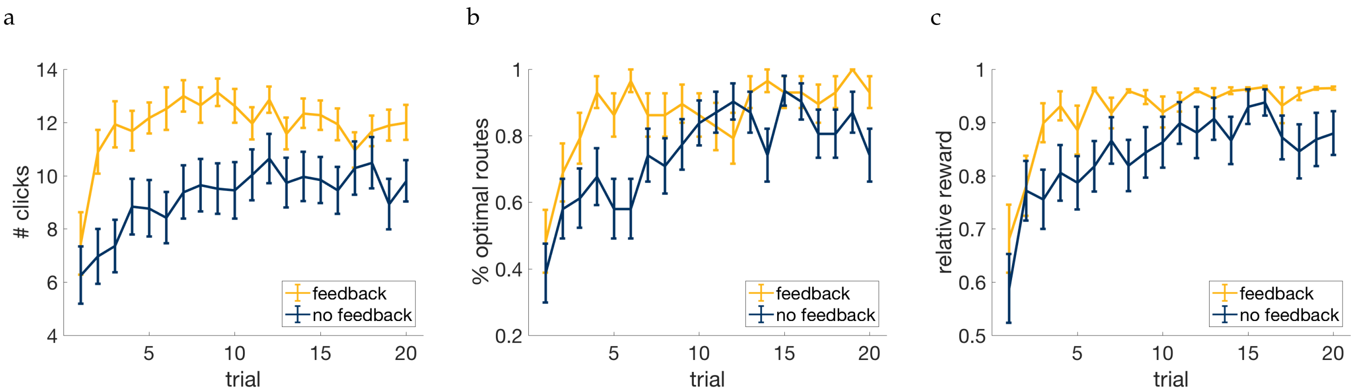

Figure 2: Learning curves for the number of outcomes considered during planning (a), the frequency of identifying the optimal path (b), and relative reward earned (c) with versus without feedback.

route ($t(1197) = 5.35, p < .0001$), and the relative reward ($t(1197) = 5.67, p < 0.0001$) in both conditions. Critically, feedback significantly increased the slope of the sigmoidal learning curves of the proportion of inspected cells ($t(1197) = 2.02, p < .05$), relative performance ($t(1197) = 3.43, p < .001$), and the relative frequency of the optimal route ($t(1197) = 3.29, p < .001$). Model comparisons confirmed that the full model including the feedback regressor explained the data significantly better than the restricted model without it (clicks: $F(1, 1196) = 95.39, p < 10^{-15}$, frequency of optimal solution: $F(1, 1196) = 38.90, p < 10^{-9}$, relative reward: $F(1, 1196) = 33.12, p < 10^{-7}$). This indicates that the feedback we generated from the meta-level pseudo-rewards significantly accelerated the process by which people learned to plan more and better. As a result, participants in the feedback condition learned to consider more possible outcomes ($11.9/16$ vs. $9.1/16$, $F(1, 1160) = 100.38, p < 0.0001$) and consequently earned more relative reward than participants in the control condition (92% vs. 84%, $F(1, 1160) = 43.03, p < .0001$).

4 Conclusion

We have derived a computational method for designing feedback structures that promote metacognitive learning and illustrated its potential to help people learn to plan better. The results presented here build on recent work showing that cognitive plasticity can be understood in terms of meta-level reinforcement learning (Krueger et al., 2017; Lieder, Shenhav, Musslick, & Griffiths, in revision). Together with the preliminary results presented here these findings inspire hope that theory and tools of reinforcement can be leveraged to improve the human mind by promoting metacognitive learning. The results presented here are a first step in evaluating meta-level reward shaping as a theoretical foundation for cognitive training. Future work will rigorously evaluate this feedback mechanism against alternatives, assess retention, and evaluate transfer. If successful this line of research might lead to principled and more effective tools and interventions for training high-level cognitive functions such as planning and decision making, cognitive control, reasoning, problem solving, attention, and learning.

References

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., … Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature*, *501*(7465), 97–101.

Bavelier, D., Green, C. S., Pouget, A., & Schrater, P. (2012). Brain plasticity through the life span: learning to learn and action video games. *Annual review of neuroscience*, *35*, 391–416.

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*.

Green, C. S., & Bavelier, D. (2008). Exercising your brain: a review of human brain plasticity and training-induced learning. *Psychology and aging*, *23*(4), 692.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Hay, N., Russell, S., Tolpin, D., & Shimony, S. (2012). Selecting computations: Theory and applications. In N. de Freitas & K. Murphy (Eds.), *Uncertainty in artificial intelligence: Proceedings of the twenty-eighth conference*. Corvallis, OR: AUAI Press.

Krueger, P. M., Lieder, F., & Griffiths, T. L. (2017). Enhancing metacognitive reinforcement learning using reward structures and feedback. In *Proceedings of the 39th annual conference of the cognitive science society*.

Lieder, F., & Griffiths, T. L. (2016). Helping people make better decisions using optimal gamification. In *Proceedings of the 38th annual meeting of the cognitive science society* (pp. 2075–80). Austin, TX: Cognitive Science Society.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. (in revision). Rational metareasoning and the plasticity of cognitive control.

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? the promise and challenges of enhancing cognition by training working memory. *Psychonomic bulletin & review*, *18*(1), 46–60.

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In I. Bratko & S. Dzeroski (Eds.), *Proceedings of the 16th Annual International Conference on Machine Learning* (pp. 278–287). San Francisco: Morgan Kaufmann.

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., … Ballard, C. G. (2010). Putting brain training to the test. *Nature*, *465*(7299), 775–778.

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

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT press.