Enhancing metacognitive reinforcement learning using reward structures and feedback

Paul M. Krueger[1] (pmk@berkeley.edu)
Falk Lieder[1] (falk.lieder@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California Berkeley, Berkeley, CA 94720 USA
[1] These authors contributed equally.

Abstract

How do we learn to think better, and what can we do to promote such metacognitive learning? Here, we propose that cognitive growth proceeds through metacognitive reinforcement learning. We apply this theory to model how people learn how far to plan ahead and test its predictions about the speed of metacognitive learning in two experiments. In the first experiment, we find that our model can discern a reward structure that promotes metacognitive reinforcement learning from one that hinders it. In the second experiment, we show that our model can be used to design a feedback mechanism that enhances metacognitive reinforcement learning in an environment that hinders learning. Our results suggest that modeling metacognitive learning is a promising step towards promoting cognitive growth.

Keywords: Decision-Making; Planning; Metacognitive Reinforcement Learning; Cognitive Training

Introduction

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 (C. S. 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.

To address these problems, we model cognitive plasticity as metacognitive reinforcement learning. 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.

Here, we evaluate this approach in the domain of planning. As a first step, we developed a metacognitive reinforcement learning model of how people learn how many steps to plan ahead in sequential decision problems, and we test its predictions empirically. The results of our first experiment suggest that our model can discern which reward structures are more conducive to metacognitive learning. In our second experiment, we find that feedback structures designed based on our model can accelerate learning to plan.

We start by introducing the theory of reinforcement learning that our approach is based upon. The following two sections apply this theory to model the problem of deciding how to decide and the process by which people learn to do so. We then use this theory to motivate a novel computational method for designing feedback structures that promote cognitive plasticity and experimentally test the predictions of our theory. We close with a discussion of the implications of our results for cognitive training.

Planning and reinforcement learning

A sequential decision problem can be modeled as a *Markov decision process* (MDP)

$$M = (\mathcal{S}, \mathcal{A}, T, \gamma, r, P_0), \tag{1}$$

where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $T(s,a,s')$ is the probability that the agent will transition from state s to state s' if it takes action a, $0 \leq \gamma \leq 1$ is the discount factor on future rewards, $r(s,a,s')$ is the reward generated by this transition, and P_0 is the probability distribution of the initial state S_0 (Sutton & Barto, 1998). A *policy* $\pi : \mathcal{S} \mapsto \mathcal{A}$ specifies which action to take in each of the states. The expected sum of discounted rewards that a policy π will generate in the MDP M starting from a state s is known as its *value function*

$$V_M^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(S_t, \pi(S_t), S_{t+1})\right]. \tag{2}$$

The optimal policy $\pi_M^\star$ maximizes the expected sum of discounted rewards, that is

$$\pi_M^\star = \arg\max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(S_t, \pi(S_t), S_{t+1})\right]. \tag{3}$$

Solving large planning problems is often intractable because the number of possible action sequences grows exponentially with the number of steps one plans ahead. When the state space $\mathcal{S}$ is discrete and relatively small, dynamic programming can be used to find optimal plans in polynomial time (Littman, Dean, & Kaelbling, 1995). But the high-dimensional, continuous state spaces people have to plan with in real life are too large for these methods. Instead, people seem to rely on approximate planning strategies (Huys et al., 2015) and often decide primarily based on immediate

and proximal outcomes while neglecting the long-term consequences of their actions (Myerson & Green, 1995). Despite its fallibility, looking only a few steps ahead can drastically simplify the planning problem, and this may often be a necessity for bounded agents with imperfect knowledge of the environment (Jiang, Kulesza, Singh, & Lewis, 2015). Since cutting corners in the decision process is both necessary and problematic, good decision-making requires knowing when that is admissible and when it is not. Knowing how much to plan is therefore an important metacognitive skill to learn.

Previous work suggests that this metacognitive skill can be learned through trial and error (Lieder & Griffiths, 2015). Learning through trial and error can be understood in terms of *reinforcement learning* (Sutton & Barto, 1998). While certain reinforcement learning algorithms can, in principle, learn to solve arbitrarily complex problems, reinforcement learning can also be very slow—especially when rewards are sparse and the optimal policy is far from the learner's initial strategy. A common approach to remedy this problem is to give the algorithm pseudo-rewards for actions that do not achieve the goal but lead in the right direction (Ng et al., 1999).While previous work has developed this idea to accelerate learning a direct mapping from states to actions, we will leverage it to accelerate learning to plan.

Deciding how to decide

People can use many different decision strategies. This poses the problem of deciding how to decide (Boureau, Sokol-Hessner, & Daw, 2015). Previous research on meta-decision-making has focused on the arbitration between habits versus planning (Keramati, Dezfouli, & Piray, 2011; Dolan & Dayan, 2013). While this is an important meta-control problem, it is only one part of the puzzle because people are equipped with more than one goal-directed decision-mechanism. Hence, when the model-based system is in charge, it has to be determined how many steps it should plan ahead. Ideally, the chosen planning horizon should achieve the optimal tradeoff between expected decision quality versus decision time (Vul, Goodman, Griffiths, & Tenenbaum, 2014) and mental effort (Shenhav et al., 2017).

Here, we make the simplifying assumption that people always choose the action that maximizes their sum of expected rewards over the next h steps, for some value of h that differs across decisions. A planning horizon of $h = 1$ entails looking only at the immediate outcome of each action (myopic one-step planning) whereas a planning horizon larger than one entails solving a sequential decision problem to form a multi-step plan. Under this assumption, the meta-decision problem is to select a planning horizon h from a set $\mathcal{H} = \{1, 2, \cdots, \}$, execute the plan, select a new planning horizon, and so on. More formally, this problem can be formalized as a meta-level MDP (Hay, Russell, Tolpin, & Shimony, 2012). In our task, the meta-level MDP is

$$M_{\text{meta}} = (\mathcal{S}_{\text{meta}}, \mathcal{H}, T_{\text{meta}}, r_{\text{meta}}), \tag{4}$$

where the meta-level state $m \in \mathcal{S}_{\text{meta}} = \{0, 1, 2, 3, 4\}$ encodes the number of remaining moves, and the meta-level action $h \in \mathcal{H} = \{1, 2, 3, 4\}$ is the planning horizon used to make a decision. The meta-level reward function r_{meta} integrates the cost of planning with the return of the resulting action:

$$r_{\text{meta}}(m_k, h_k) = -\text{cost}(h_k) + \sum_{t=1}^{h} r(s_t, \text{plan}_t^{(k,h_k)}), \tag{5}$$

where $\text{plan}_t^{(k,h)}$ is the t^{th} action of the plan formed by looking h steps ahead in the meta-level state m_k. The meta-decision-maker receives this reward after the plan has been executed in its entirety. If the meta-decision-maker selects short planning horizons there can be multiple plan-act-reward-learn cycles within a single trial. The cost of planning $\text{cost}(h_k)$ is determined by the branching factor b of the decision tree according to

$$\text{cost}(h_k) = \lambda \cdot b^{h_k} \cdot h_k, \tag{6}$$

where b^{h_k} is the number of plans, h_k is the number of steps per plan, and λ is the cost per planning step.*

Metacognitive reinforcement learning

Solving the problem of deciding how to decide optimally is computationally intractable but the optimal solution can be approximated through learning (Russell & Wefald, 1991). We propose that people use reinforcement learning (Sutton & Barto, 1998) to approximate the optimal solution to the meta-decision problem formulated in Equation 4.

Model

Our model of metacognitive reinforcement learning builds on the semi-gradient SARSA algorithm (Sutton & Barto, 1998) that was develop to approximately solve MDPs with large or continuous state spaces. Specifically, we assume that people learn a linear approximation to the meta-level Q-function

$$Q_{\text{meta}}(m_k, h_k) \approx \sum_{j=1}^{7} w_j \cdot f_j(m_k, h_k), \tag{7}$$

whose features $\mathbf{f}$ comprise one indicator variable for each possible planning horizon h ($f_1 = \mathbb{1}(h = 1), \cdots, f_4 = \mathbb{1}(h = 4)$), one indicator variable for whether or not the agent planned all l steps until the end of the task ($f_5 = \mathbb{1}(h = l)$), the number of steps that were left unplanned ($f_6 = \max\{0, l - h\}$), and the number of steps the agent planned too far ($f_7 = \max\{0, h - l\}$). The semi-gradient SARSA algorithm learns the weights of these features by gradient descent. To bring it closer to human performance, our model replaces its gradient descent updates by Bayesian learning. Concretely, the weights $\mathbf{w}$ are learned by Bayesian linear regression of the bootstrap estimate $\hat{Q}(m_k, h_k)$ of the meta-level value function onto the features $\mathbf{f}$. The bootstrap estimator

$$\hat{Q}(m_k, h_k) = r_{\text{meta}}(m_k, h_k) + \langle \mu_t, \mathbf{f}(m', h') \rangle \tag{8}$$

*This equation assumes a constant branching factor and an upper bound on the complexity of planning. People's planning time likely increases less than exponentially fast with the planning horizon but our approximation may be sufficient for small problems.

is the sum of the immediate meta-level reward and the predicted value of the next meta-level state m'. The predicted value of m' is the scalar product of the the posterior mean μ_t of the weights $\mathbf{w}$ given the observations from the first t actions (where $t = \sum_{n=1}^{k} h_n$) and the features $\mathbf{f}(m', c')$ of m' and the planning horizon h' that will be selected in that state.

We assume that the prior on the feature weights reflects that it is beneficial to plan until the end ($P(f_5) = \mathcal{N}(\mu = 1, \sigma = 0.1)$), although planning is costly ($P(f_1) = P(f_2) = P(f_3) = P(f_4) = \mathcal{N}(\mu = -1, \sigma = 0.1)$), and that planning too much is more costly than planning too little ($P(f_7) = \mathcal{N}(\mu = -1, \sigma = 0.1)$ and $P(f_6) = \mathcal{N}(\mu = 0, \sigma = 0.1)$).

Given the posterior on the feature weights $\mathbf{w}$, the planning horizon h is selected by Thompson sampling. Specifically, to make the k^{th} meta-decision, a weight vector $\tilde{w}$ is sampled from the posterior distribution of the weights given the series of meta-level states, selected planning horizons, and resulting value estimates experienced so far. That is,

$$\tilde{w}_k \sim P(\mathbf{w}|\mathcal{E}_k), \tag{9}$$

where the set $\mathcal{E}_k = \{e_1, \cdots, e_k\}$ contains the meta-decision-maker's experience from the first k meta-decisions; to be precise, each meta-level experience $e_j \in \mathcal{E}_k$ is a tuple $\left(m_j, h_j, \hat{Q}(m_j, c_j; \mu_j)\right)$ containing a meta-level state, the computation selected in it, and the bootstrap estimates of its Q-value. The sampled weight vector $\tilde{w}$ is then used to predict the Q-values of each possible planning horizon $h \in \mathcal{H}$ according to Equation 7. Finally, the planning horizon with the highest predicted Q-value is used for decision-making.

By proposing metacognitive reinforcement learning as a mechanism of cognitive plasticity, our model suggests that reward and feedback are critical for cognitive growth. Conceptualizing metacognitive reinforcement learning as a regression problem suggests that learning how to best think about a problem should require less practice the stronger the correlation between the features $\mathbf{f}(m, c)$ (i.e., the predictors) and the resulting reward net the cost of thinking (i.e., the criterion; Green, 1991). Here, we apply our model to predict how quickly people can learn that more planning leads to better results from the reward structure of the practice problems. According to the model, learning should be fastest when the reward increases deterministically with the planning horizon both within and across problems. By contrast, learning should be slower when this relationship is degraded by additional variability in the rewards that is unrelated to planning. The following experiments test this prediction and illustrate the model's utility for designing feedback structures that promote metacognitive learning.

Experiment 1: Reward structures can help or hinder learning to plan

Methods

We recruited 304 adult participants from Amazon Mechanical Turk. The task took about 25 minutes, and participants were paid \$2.50 plus a performance-dependent bonus of up to

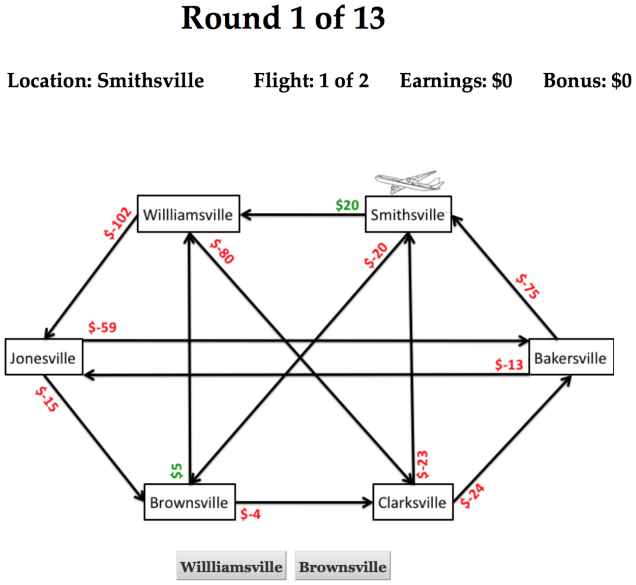

Figure 1: Screenshot of a problem from Experiment 1.

\$2.00. Participants played a series of *flight planning* games. The environment consisted of six different cities, each connected to two other cities (Figure 1). Participants began each trial at a given city, and were tasked with planning a specified number of flights. Each flight was associated with a known gain or loss of money, displayed onscreen. Thus, the participants' task was to plan a route that would maximize their earnings or minimize their losses, based on the number of planning steps required for that game.

The experiment comprised thirteen trials total: a sequence of three practice problems which required planning 2, 3, and 3 steps ahead, respectively, followed by ten 4-step problems, with a break after trial eight. The order of the two 3-step problems was randomized, and the order of the ten 4-step problems was randomized across the last ten trials of the experiment. Participants were assigned randomly to one of two conditions: environments with reward structures designed to promote learning ("diagnostic rewards"), or environments with reward structures designed to hinder learning ("non-diagnostic rewards").

The problems of the diagnostic rewards condition were automatically generated to exhibit four characteristics:

1. For each l-step problem, planning $h < l$ steps ahead generates $l - h$ suboptimal moves. In other words, each myopic planner makes the maximum possible number of mistakes.

2. When the number of moves is l, then planning l steps ahead yields a positive return, but planning $h < l$ steps ahead yields a negative return.

3. The return increases monotonically with the planning horizon from 1 to the total number of moves.

4. Each starting position occurs at least once.

The reward structures used for the non-diagnostic rewards condition were created by shifting the diagnostic reward structures so as to degrade the correlation between planning horizon and reward. Concretely, for half of the problems all rewards were shifted down such that no amount of planning could achieve a return better than $-\$10$. Since the original problems were such that the 1-step planner always performed worst, the shift was $\frac{-r_1+X}{l}$ where r_1 is the return of the 1-step planner, l is the number of steps in the planning problem, and X is a random number between 10 and 20 that differed across problems ($X \sim \text{Uniform}([10,20])$). For the other half of the problems, all rewards were shifted up by $-\frac{r_1+X}{l}$ such that all planners achieve a return of at least $+\$10$. These reward structures make it extremely difficult for metacognitive reinforcement learning to discover that planning is valuable, because the random shifts greatly diminish the correlation between planning horizon and reward.

Results

Both model simulations and human behavior demonstrated enhanced learning in environments with diagnostic rewards. Figure 2 shows the mean performance of the metacognitive reinforcement learning model, and the mean performance of human participants. Here, performance is measured as relative reward

$$R_{rel} = (R - R_{min})/(R_{max} - R_{min}), \tag{10}$$

where R is the total reward received during the trial, and R_{min} and R_{max} are the highest and lowest possible total reward on that trial, respectively.

To measure the effects of condition and trial number on performance in human participants, we ran a repeated-measures ANOVA. This revealed a significant effect of both trial number ($F(9,2989) = 3.44, p < 0.001$) and condition ($F(9,3029) = 15.26, p < 0.0001$), such that participants improved over time, and participants with diagnostic feedback performed better than those without. To measure learning in each group, we ran a simple linear regression of the relative reward on the trial number. This revealed a significant regression equation for participants who received diagnostic rewards ($F(2,302) = 11.28, p < 0.01$), with an R^2 of 0.59, but not for participants who received non-diagnostic rewards ($F(2,302) = 3.51, p > 0.05$), with an R^2 of 0.31, suggesting that improvement in performance occurred with diagnostic rewards, but not without.

To analyze the frequency with which participants chose the optimal route, we performed a multinomial logistic regression of whether or not each participant chose the optimal route on trial number and group. This revealed significant effects of trial number ($p < 10^{-6}$) and group ($p < 0.0001$).

In addition, we found that participants interacting with a diagnostic reward structure learned to plan significantly further ahead than participants interacting with the non-diagnostic reward structure. When there were four steps left, the average planning horizon was 2.96 with diagnostic rewards compared to 2.65 with non-diagnostic rewards ($t(596) = 2.94$, $p < 0.01$). When the rewards were diagnostic of good planning, participants' choices in the first step of the 4-step problems accorded 10.3% more frequently with 4-step planning ($t(302) = 3.57, p < 0.001$). For 3 remaining steps there was a significant increase in choices according with optimal 1-step ($p < 0.01$), 2-step ($p < 0.01$) and 4-step planning ($p < 0.01$). For 2 remaining steps, there was a significant increase in choices according with optimal 1-step planning ($p < 0.0001$) without a decrease in agreement with other planning horizons. Finally, on the last move participants' choices in the environment with diagnostic rewards corresponded 5.8% more frequently with optimal 1-step planning ($t(302) = 3.71, p < 0.001$), and significantly less frequently with 2-step and 3-step planning ($p < 0.01$ and $p < 0.001$). In summary, diagnostic rewards led to better agreement between the planning horizon and the number of remaining steps.

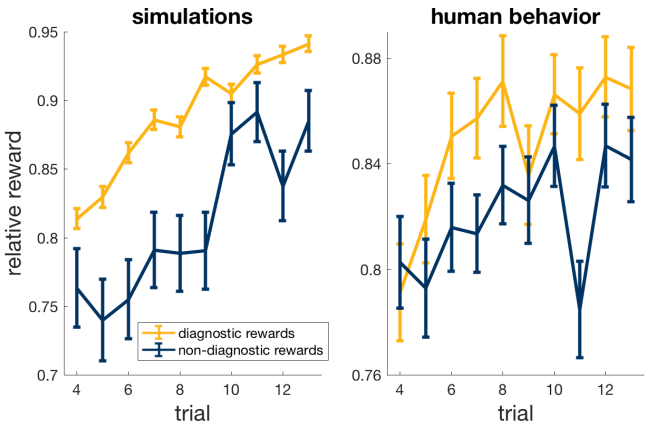

Figure 2: Model predictions and human performance in Experiment 1. Error bars indicate the standard error of the mean. Model predictions were averaged over 500 simulations.

Experiment 2: Using feedback to promote learning to plan

When one has control over the reward structure of an environment, creating rewards tailored to faster learning may be feasible. However, often environmental rewards are fixed. In Experiment 2, we tested whether providing feedback may be an effective alternative approach to accelerating learning. When participants do not plan enough to find the optimal route, this could be because the time cost of planning an optimal route outweighs its benefits. To change that, we provided feedback in the form of timeout penalties for short-sighted decisions.

Methods

We recruited 324 adult participants on Amazon Mechanical Turk. The task took about 30 minutes, and participants were paid $3.00 plus a performance-dependent bonus of up to $2.00. Participants played twenty trials of the flight planning game described above. These trials were divided into a training block and a testing block. The training block con-

sisted of six trials requiring 2-step planning, followed by ten trials requiring 3-step planning. The testing block consisted of four additional 3-step trials. The order of the 2-step trials and the order of the 3-step trials were randomized across subjects. Participants were randomly assigned to either the feedback condition or the control condition.

In the training block, participants in the feedback condition were told their apparent planning horizon at the end of every trial and penalized with a timeout that reflected the amount of planning they had eschewed. Concretely, we set the durations of the timeouts such that the cost of short-sighted decisions was proportional to the amount of necessary planning the participant had eschewed. Specifically, the forgone cost of planning was estimated by $\text{cost} = 2^{l-\hat{h}}$, where l is the number of moves for that trial, $\hat{h}$ is the participant's apparent planning horizon, and 2 is the branching factor since each step entailed a binary decision. The participant's planning horizon was estimated by the number of consecutive moves consistent with the optimal policy, beginning with the last move, followed by the second-to-last, etc. At the end of each trial of the first block, participants in the feedback group were penalized with a timeout delay for sub-optimal routes. The delay was calculated as $7 \cdot (\text{cost} - 1)$ seconds. During this period, participants were unable to proceed to the next trial. If participants performed the optimal route, they were able to proceed immediately to the next trial.

The control group received no feedback and had to wait a fixed amount of time at the end of every trial in block 1, regardless of their performance. This fixed period was set to 8 seconds, to match the mean timeout period for participants in the feedback group (7.9 seconds). Neither group received feedback or delays in the test block.

The planning problems presented in this experiment were created in two steps. In the first step, we created 2- and 3-step problems with maximally diagnostic reward structures (according to the criteria used in Experiment 1) subject to the constraint that the first move with the highest immediate reward was optimal for exactly half of those problems. In the second step, we modified these problems so as to deteriorate the correlation between planning horizon and reward using the same method we employed to create the non-diagnostic reward structures used in Experiment 1.

Model Predictions

We applied the metacognitive reinforcement learning model described above to the problem of learning how many steps one should plan ahead. We simulated a run of the experiment described above with 1000 participants in each condition. The simulations predicted a gradual increase in the relative return from the first 3-step problem to the last one (see Figure 3). With feedback, the relative return increased faster and reached a higher level than without feedback.

Results

To quantify the effects of condition and trial number on performance (measured as relative reward), we ran a mixed-design repeated-measures ANOVA on participant performance during the 3-step trials. This revealed a significant effect of feedback ($F(9,4521) = 8.54, p < 0.01$) and trial number ($F(9,4521) = 1.85, p < 0.05$) on relative reward. To measure learning in each group, we performed a simple linear regression of relative reward on trial number for the 3-step trials in the training block (i.e., when participants in the feedback group received feedback). This revealed a significant regression equation for the feedback group ($F(2,322) = 5.28, p = 0.05$), with an R^2 of 0.40 but not for the control group ($F(2,322) = 1.57, p > 0.05$), with an R^2 of 0.16. This suggests that participants who received feedback improved during the training block but the control group did not.

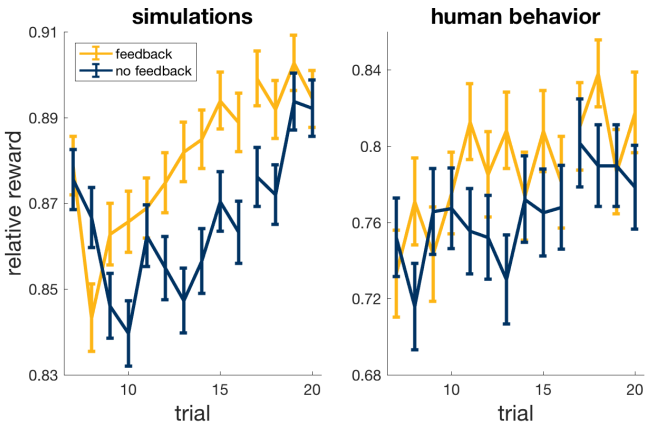

Figure 3: Results of Experiment 2. The metacognitive RL model predicts that feedback accelerate learning to plan. Human behavior shows a similar pattern of results.

Feedback increased the model's average performance in both the training block and the transfer block. We next tested whether the enhanced learning of the feedback group during training resulted in better performance in the transfer block (trials 17-20) where they no longer received any feedback. A two-sample t-test revealed that the feedback group's advantage in the testing block was nearly significant ($t(1294) = 1.53, p = 0.063$). Figure 3 compares our participants' performance to the model predictions.

As predicted by our model, a multinomial logistic regression of whether or not each participant chose the optimal route on trial number and feedback, revealed significant effects of trial number ($p < 0.0001$) and feedback ($p < 0.01$).

Feedback appeared to increase people's planning horizons: when there were two remaining moves, the choices of the feedback group accorded 4% less often with myopic choice ($t(1398) = -2.17, p < 0.05$), 7% more often with optimal 2-step planning ($t(1398) = 3.44, p < 0.001$), and 4% more often with optimal 3-step planning ($t(1398) = 2.43, p < 0.05$).

Discussion

In this article, we have introduced a computational model of how people learn to decide better. Its central idea is that

learning how to think can be understood as metacognitive reinforcement learning. Our model extends previous research on strategy selection learning (Lieder et al., 2014; Lieder & Griffiths, 2015) by capturing that choosing cognitive operations is a sequential decision problem with potentially delayed rewards rather than a one-shot decision. The new model correctly predicted the effects of reward structure and feedback on learning to plan: Experiment 1 suggested that our model captures the effect of reward structures on the speed of metacognitive learning. We then applied our theory to design feedback for people's performance in environments whose reward structure is not diagnostic of good planning. Experiment 2 confirmed the model's prediction that this intervention would be effective.

Our results suggest two pragmatic approaches to promoting cognitive growth: first, designing reward structures that are diagnostic of the quality of reasoning, planning, and decision-making; second, providing feedback on the process by which a decision was made. In Experiment 2 we followed the latter approach by designing feedback based on the cost of planning; but other types of feedback may also be useful. If cognitive plasticity is based on model-free reinforcement learning as assumed by our theory, then its speed should critically depend on how well the feedback people receive upon performing cognitive operations reflects their value. Therefore, feedback structures that align immediate feedback with long-term value should be maximally effective at promoting cognitive plasticity and learning to make better decisions. Future experiments should test this hypothesis by designing feedback structures using the optimal gamification method introduced by Lieder and Griffiths (2016). Feedback designed using optimal gamification could be especially beneficial because the underlying method of reward shaping is designed to accelerate model-free reinforcement learning (Ng et al., 1999). Critically, to promote learning how to decide, people should decide without any assistance and only receive feedback *after* their choice.

We hope that our theory of metacognitive reinforcement learning will be a step towards establishing a scientific foundation for designing feedback for cognitive training and other interventions for promoting cognitive growth. Future work will evaluate alternative forms of feedback, address the problem of transfer and retention, and design more effective training paradigms using tasks that are maximally diagnostic of how people think and decide.

Acknowledgments. This work was supported by grant number ONR MURI N00014-13-1-0341. The authors thank Thomas Hills, Peter Dayan, Rika Antonova, Silvia Bunge, Stuart Russell, Amitai Shenhav, and Sebastian Musslick for feedback and discussions.

References

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., ... others (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.

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.

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

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.

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate behavioral research*, *26*(3), 499–510.

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.* P.O. Box 866 Corvallis, Oregon 97339 USA: AUAI Press.

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098–3103.

Jiang, N., Kulesza, A., Singh, S., & Lewis, R. (2015). The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems* (pp. 1181–1189).

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput Biol*, *7*(5), e1002055.

Lieder, F., & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In *Proceedings of the 37th annual conference of the cognitive science society.*

Lieder, F., & Griffiths, T. L. (2016). Helping people make better decisions using optimal gamification. In *Proc. 38th annu. conf. cogn. sci. soc., philadelphia* (pp. 2075–80).

Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2870–2878). Curran Associates, Inc.

Littman, M. L., Dean, T. L., & Kaelbling, L. P. (1995). On the complexity of solving markov decision problems. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 394–402).

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.

Myerson, J., & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the experimental analysis of behavior*, *64*(3), 263–276.

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.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T., Cohen, J., & Botvinick, M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*.

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

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.