Learning planning strategies without feedback

Srinidhi Srinivas (srinidhi.srinivas@tuebingen.mpg.de)

Ruiqi He (ruiqi.he@tuebingen.mpg.de)

Falk Lieder (falk.lieder@tuebingen.mpg.de)
Max Planck Institute for Intelligent Systems
Max Planck Ring 4
Tübingen, 72076 DE

Abstract

How do humans get better at planning? Previous work postulated that the improvement of cognitive strategies occurs through feedback-based *metacognitive reinforcement learning* (MCRL). However, it is not clear whether and, if so, how people can learn planning strategies without reinforcement. To answer these questions, we experimentally investigated the effect of frequency of feedback on people's ability to learn adaptive planning strategies. We found that participants receiving feedback only 25% of the time nonetheless learned about as well as participants receiving constant feedback. Quantitative modelling of the data revealed that state-of-the-art MCRL models cannot explain this finding. However, extending these models by a mechanism generating an additional learning signal through self-evaluation of plan quality can account for people's ability to learn planning strategies without feedback. The findings of this research have implications for the design of learning environments and enabling people and machines to self-sufficiently improve their strategies in naturalistic settings.

Keywords: metacognitive learning; strategy learning; cognitive skill acquisition; reinforcement learning; decision-making

Introduction

People have been shown to use adaptive cognitive strategies across diverse scenarios (Anderson, 2013; Lieder & Griffiths, 2020; Callaway et al., 2022). By employing a variety of strategies, humans can meet the demands of environments that differ vastly in structure and constraints. But people are not endowed with all the concrete strategies required to overcome specific challenges. Instead, some of these cognitive skills are acquired through learning from trial and error (VanLehn, 1996; Siegler, 1999; Siegler & Jenkins, 2014).

One such mechanism is learning from feedback (Brinko, 1993). An obvious example is the feedback of an experienced teacher, who evaluates the appropriateness of a student's strategies. For example, a medical student, tasked with diagnosing a patient based on certain symptoms, receives a score about the quality of their diagnosis, and likely also explicit feedback on improving their process of investigation.

Recent work on *metacognitive learning*, the cognitive process of adaptively improving cognitive skills, supports the presence of feedback-driven reinforcement learning mechanisms in the human brain for this type of adaptive learning (Krueger, Lieder, & Griffiths, 2017; Lieder, Shenhav, Musslick, & Griffiths, 2018; Jain et al., 2019; He & Lieder, 2022). Computational models of this type of learning postulate multiple ways in which the brain represents planning strategies, and how the brain uses feedback from the environment to iteratively adjust these strategies to improve performance.

However, it is not clear that metacognitive learning from feedback is sufficient to capture the robustness of people's ability to learn adaptive cognitive strategies. Most real-world scenarios do not provide explicit feedback on people's actions, let alone their cognitive strategies. For example, a student practicing writing timed essays for a standardized test may not get detailed feedback on each attempt, but can still consistently improve the quality and the efficiency of their writing.

The existence of this latter type of learning motivates the present study. Existing work within the paradigm of *metacognitive reinforcement learning* (MCRL) does not address the question of whether and, if so, how people learn planning strategies in the absence of explicit feedback.

To answer this question, we investigated how people learn planning strategies from trial and error. We administered several rounds of a game that requires planning, while varying whether and how often participants receive feedback about their plan quality on each round. Our experiment revealed that participants improve their strategies comparably well in both feedback-scarce and feedback-rich environments. To quantitatively capture this phenomenon, we extended upon the state-of-the-art MCRL models to develop the first MCRL models that generate self-evaluation signals to learn in the absence of feedback. Finally, we leveraged the experiment data to quantitatively compare these extended models against the previous state of the art, and answer the question of how participants might have been able to learn better planning strategies. This brings us one step closer to answering important practical questions, such as "How can we teach human and computational agents to self-sufficiently improve their strategies in more naturalistic settings?"

The remainder of the paper is structured as follows: the next section details the experiment and its results. The section after describes the computational models tested to explain the learning mechanisms of the participants, how those models were quantitatively compared, and the results of those comparisons. Finally, the results from the experiment and the comparison of models are discussed in light of the bigger question addressed by this study - how do people adaptively acquire cognitive skills?

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

Experiment

Methods

Participants 208 participants were recruited for the online experiment through the recruitment platform Prolific. The ages of the participants ranged from 17 to 66, with the average participant age being 27.8 (SD = 8.7). Of the 208 participants, 125 identified as male, 82 as female, and 1 as 'other.'

Participants were paid a base amount of £4.50 for completing the full experiment. In addition, they were eligible for a bonus of up to £2 proportional to their final score at the end of the 3-step planning task. The average bonus earned was £1.39 (SD = 0.54)

Experimental Design Participants were randomly assigned to be either in the feedback-scarce (scarcity) condition (N=105) or in the feedback-rich (control) condition (N=103). In both of the conditions, the main task was to complete several trials of slightly different versions of the *3-step planning task*, a sequential planning task which is described later in detail.

Materials To trace participants' planning strategies during each trial of the experiment, the planning task was administered using the *Mouselab-MDP* paradigm (Callaway, Lieder, Krueger, & Griffiths, 2017). Mouselab-MDP is a paradigm for sequential planning tasks that represents people's beliefs and planning operations in the form of states and actions in the environment. People's interaction with the environment externalizes their planning behavior, and is therefore diagnostic of the planning strategy employed on each iteration of the task (Callaway et al., 2017, 2022; Jain et al., 2022).

Each trial of the standard 3-step planning task involves navigating a spider through a web consisting of nodes and edges that are arranged as a directed graph of depth 3. Each of the nodes contains an integer monetary value.

At the beginning of a trial, all the node values are obscured. The participant can click any number of nodes to reveal their values in exchange for a fixed cost of $1 that is deducted from the participant's trial score, for each click made. This clicking action corresponds to the planning operation of collecting information about a future state, and thus incurs a cognitive cost, such as when a doctor uses effort and resources to conduct a certain medical test when devising a treatment plan for a patient's symptoms. Participants are therefore encouraged to use their clicks strategically.

After revealing the node values, the participant moves the spider from the starting node to a leaf node of the web in 3 steps. The participant may not click to reveal information about a node after the spider has begun moving. Whenever the spider passes through a node, the value present at that node is added to the participant's score for that trial. This step represents the execution of the participant's plan, and the score they receive at the end of the trial is indicative of the quality of their employed strategy. This sequence is then repeated for several trials, so that participants have an opportunity to improve their planning strategy in structurally similar

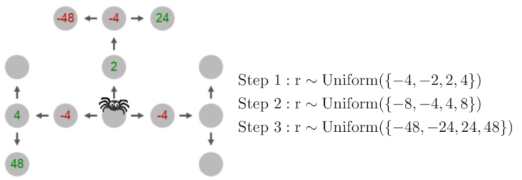

Figure 1: The 3-step planning task implemented in the Mouselab-MDP paradigm. Distributions of the node rewards at every level are shown on the right.

environments. The participant's total score for the experiment is the sum of the scores obtained on all the individual trials (including deducted costs.) The task of the participant is to maximize the total score obtained while minimizing the total costs incurred over all the trials of the experiment. The participant's total experiment score is shown to them only after completion of all the trials.

On every trial, the values at a node are sampled uniformly from a distribution of 4 different values, centered around 0. Sampling rewards ensures structural similarity of the environment across trials while varying the rewards presented to the participant. For the nodes at the first level (one step away from the starting node), the distribution has the lowest variance, with the lowest and highest possible values being -$4 and +$4 respectively. For the nodes at the second level, the extreme values of the distribution were ±$24, and for the nodes at the third level, ±$48 (see Figure 1). These reward distributions, where the most valuable nodes are the furthest ones, were chosen to encourage learning of far-sighted planning behavior.

Procedure In both conditions, participants were given instructions and completed two example trials of the 3-step planning task. Then, everyone had 3 attempts to achieve a perfect score on a 6-question quiz about the task instructions before they were allowed to proceed to the actual task.

In the scarcity condition, participants completed 120 trials of a slightly modified 3-step planning task over the course of four equally sized blocks. In this modified version of the task, the trial score was displayed at the end of only 25% of the trials (30 trials), distributed evenly over the blocks. On the remaining 90 trials, the participants were told, at the end of the round, that the spider forgot to count how much money was collected on that trial, and that the trial score was therefore unknown. Until the end of the trial, it was not made known to the participant whether they would receive a score on that trial or not. Nonetheless, costs for each of the clicks were still deducted on every trial. To compensate for the lack of rewards contributing to their overall score for the experiment, the cost of each click was proportionately decreased to $0.25. This ensured that the expected experiment score and the optimal click strategy on a given trial remained constant across both conditions.

In the control condition, participants completed 30 trials of

the standard 3-step planning task described above, in a single block. To ensure that the overall experiment duration was the same across both conditions, participants in the control condition were additionally given two sets of trials of the Stroop task (Stroop, 1935), which served as task-irrelevant filler trials. Each set consisted of 450 trials grouped into 5 blocks of 90 trials each. The first set was completed before the 3-step planning task, and the second set was completed after it. The participant's performance on these filler trials did not contribute to their total score.

Data Analysis The pre-registration for the statistical analysis of participant data is available at https://osf.io/n6x5k/.

The outcome measure *expected trial score* was calculated as the expectation of the total rewards obtained along the best path from start to end, given the information that was revealed, minus the costs incurred by clicking to reveal that information.

We investigated the temporal evolution of the average outcome measures in each condition, and found that both groups exhibited a steep increase in performance in the first few trials, while slowing down over the remainder of the experiment. To make the data adhere to the linearity assumption of our Linear Mixed Model (LMM) regression analyses, we divided each condition into two phases: the *fast learning phase* and the *slow improvement phase*. We identified the best partitions for these phases by finding a general linear model of best fit that captures two linear relationships in each dataset. We thus defined the *fast learning phase* as lasting from trials 1-9 and 1-13, and the *slow improvement phase* as lasting from trials 10-30 and 14-120, in the control and scarcity conditions respectively (Figure 2).

Results

Increase in Performance Over Time Our analysis of the increase in the expected trial score over time revealed that the scores of participants in both groups improved significantly over time in both the fast learning phase ($\beta = 1.52$; $p < .001$) and slow improvement phase ($\beta = 0.22$; $p < .001$). Despite less feedback in the scarcity condition, there was no significant difference between the two conditions in the increase in performance over time in the fast learning phase ($\beta = -0.319$; $p = 0.119$). However, during the slow improvement phase, participants' performance improved significantly faster in the control condition than in the scarcity condition ($\beta = -0.172$; $p < .001$).

Effect of Feedback versus No Feedback We assessed the effect of number of previous trials with feedback and number of previous trials without feedback on the increase in outcome measures in each phase using linear mixed models. We found that, in the fast learning phase, both trials with feedback ($\beta = 1.515$; $p < .001$) and without feedback ($\beta = 1.143$; $p < .001$) significantly increased participants performance in both conditions. Participants in the scarcity condition did not learn significantly differently from trials with feedback compared to participants in the control condition ($\beta = -0.179$; $p = 0.73$). Furthermore, it was not evident that participants in either condition learned more from trials with feedback than from trials without feedback or vice versa. According to contrast tests, the difference between the corresponding LMM coefficients were not significantly different (see Table 1).

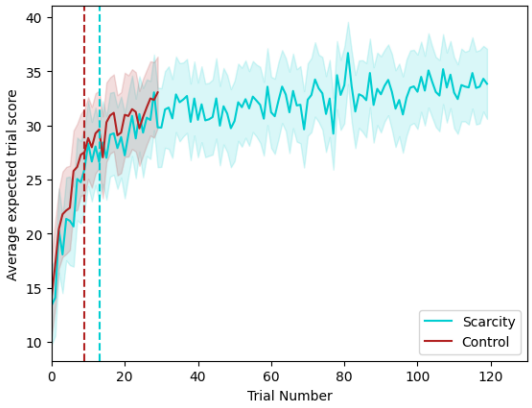

Figure 2: Improvement of long-term expected trial score over time, with 95% CI. Partitions between *fast learning phase* and *slow improvement phase* in each group marked by dotted lines.

Table 1: Results of t-contrast tests for differences between effects. Significant effects are marked in bold.

No.	Null Hypothesis
1	Feedback - No Feedback
2	Feedback + Feedback × Scarce - No Feedback

Phase	Hypothesis No.	T-test β	T-test p
Fast Learning	1	0.3720	0.163
	2	0.1935	0.757
Slow Improvement	1	**0.2623**	< 0.001
	2	**0.3662**	0.006

In the slow improvement phase, on the other hand, participants in both conditions learned only from trials with feedback ($\beta = 0.217$; $p < .001$). The effect of learning from trials with feedback was furthermore significantly greater than the effect of learning from trials without feedback, both within and between conditions (Table 1).

Difference in Experiment Performance Surprisingly, we found that the participants in the scarcity condition were able to achieve a similar level of performance in the full experi-

ment as participants in the control condition. The distributions of the participants' average expected scores in the slow improvement phases of the scarcity and the control conditions had means 32.01 ($SD = 12.38$) and 30.27 ($SD = 13.65$), respectively. These distributions were not significantly different (Mann–Whitney $U = 5230.0$, $P = 0.662$, two-tailed).

Modelling

Methods

Motivated by the results reported in the previous section, we pursued the question of which specific learning mechanism best explains the pattern of learning in feedback-scarce environments observed in participants. Particularly, we aimed to explain the finding that people can improve their strategies even after trials without explicit feedback. To this end, we selected state-of-the-art metacognitive learning models, and further added extensions that allow these models to learn in the absence of feedback. We then performed a series of systematic Bayesian model comparisons (Raftery, 1995) to find the most plausible learning mechanism underlying human performance.

MCRL Models Metacognitive reinforcement learning (MCRL) models postulate that the brain approaches the task of metacognitive learning as finding the optimal solution to a *meta-level Markov Decision Process* (Griffiths et al., 2019):

$$M_{\text{meta}} = (\mathcal{B}, \mathcal{C} \cup \{\perp\}, T_{\text{meta}}, r_{\text{meta}}), \tag{1}$$

where belief states $b_t \in \mathcal{B}$ represent the participant's beliefs about the values of different courses of actions. The belief states $(b_1, b_2 ... b_t, b_{t+1})$ temporally evolve according to the meta-level transition probabilities $T_{\text{meta}}(b_t, c_t, b_{t+1})$. Each meta-level action $c_t \in C$ corresponds to a planning operation, with the additional action $c_t = \perp$ being the decision to terminate planning. Finally, $r_{\text{meta}}(b_t, c_t)$ encodes either the cost of performing the planning operation c_t while in belief state b_t, or the expected reward for terminating planning, if $c_t = \perp$.

REINFORCE Models MCRL models approximate the optimal solution to the meta-level MDP using reinforcement learning (Krueger et al., 2017). The types of models we exclusively considered, REINFORCE models, do this by assuming that people directly adjust their planning strategy by using gradient ascent to learn a soft-max policy, which selects actions based on weighted combinations of a set of 56 neuroscience-informed features of planning strategies (He, Jain, & Lieder, 2021). Concretely, the policy is defined as:

$$\pi_\theta(c|b) = \frac{\exp(\frac{1}{\tau} \cdot \sum_{k=1}^{56} \theta_k \cdot f_k(b,c))}{\sum_{c \in C_b} \exp(\frac{1}{\tau} \cdot \sum_{k=1}^{56} \theta_k \cdot f_k(b,c))} \tag{2}$$

where $c \in C_b$ is the action being considered out of all possible actions in belief state b, f_k are the values of the strategy feature for that action in that belief state. The weights θ_k of the features, representing the planning strategy of the brain, are updated each trial according to the learning rule:

$$\theta \longleftarrow \theta + \alpha \cdot \sum_{t=1}^{O} \gamma^{t-1} \cdot r_{\text{meta}}(b_t, c_t) \cdot \nabla_\theta \ln \pi_\theta(c_t|b_t) \tag{3}$$

where α is the learning rate, γ is the discount factor, and O is the number of planning operations performed on that trial (number of clicks $c_1, ..., c_{O-1} \in C$ plus one for $c_O = \perp$). Both α and γ are treated as free parameters and fit separately for each participant.

Extended Models In environments where feedback is present, the final meta-level reward, $r_{\text{meta}}(b_t, c_t = \perp)$, is equal to the reward collected from traversing the chosen path, R_{env}. In feedback-scarce environments, however, this environmental feedback is often non-existent, causing the models to perform fewer parameter updates and learn less. To explain how people learn in the absence of feedback, we considered several extensions of the REINFORCE models that can make appropriate parameter updates on non-feedback trials. Concretely, we postulate two ways in which people might learn without feedback: pseudo-rewards, which indicate the value of information obtained after each individual planning operation (click); and a single, self-generated meta-reward at the end of a trial either in the absence of, (or in addition to) R_{env}.

The pseudo-reward (He et al., 2021) of performing a certain click c_t in belief state b_t and transitioning to the belief state b_{t+1} measures the improvement in the model's policy due to the information revealed by c_t. It is computed as the difference between the expected returns of the paths favored in the new belief state (b_{t+1}) versus the previous belief state (b_t). The expected returns are computed with respect to the new belief state. Pseudo-rewards are used to update the model parameters according to Equation 3 after each click. Concretely, the pseudo-reward is

$$PR(b_t, c_t, b_{t+1}) = \mathbb{E}[R_{\pi_{b_{t+1}}} | b_{t+1}] - \mathbb{E}[R_{\pi_{b_t}} | b_{t+1}]. \tag{4}$$

The self-evaluation, R_{self}, on the other hand, is the feedback that a rational agent would expect to receive at the end of the trial after the plan has been executed. Intended to represent the self-assessed quality of the plan, the self-evaluation was computed as the expected value of the total rewards collected from traversing the best path, evaluated based on the final belief state. The final meta-level reward for terminating planning, $r_{\text{meta}}(b_t, c_t = \perp)$ (abbreviated as R_{fm}), was then re-defined using this self-evaluation in the following way:

$$R_{\text{fm}} = \begin{cases} \beta \cdot R_{\text{env}} + (1-\beta) \cdot R_{\text{self}}, & \text{if } R_{\text{env}} \text{ present} \\ R_{\text{self}}, & \text{if } R_{\text{env}} \text{ absent,} \end{cases} \tag{5}$$

where β is the weight the model assigns to feedback when it is present. Different choices for β lead to different models (see Table 2). Models with $\beta = 0$ rely exclusively on self-evaluation, even in the presence of explicit feedback. Models with $\beta = 1$ use self-evaluations to update parameters only in the absence of explicit feedback. In addition, we also tested

Table 2: Table of considered model families and key features: how the environmental feedback is prioritized in the final meta-reward and how the model handles missing feedback (feedback-scarce trial). Type "Vanilla" represents the previous state-of-the-art. The three other types are extensions that use self-evaluation. Bold letters represent the abbreviated family name.

Model Type	β	R_{env} absent?
"**V**anilla"	fixed at 1	No parameter update
"**P**rioritize **S**elf"	fixed at 0	$R_{\text{fm}} = R_{\text{self}}$
"**P**rioritize **E**nv"	fixed at 1	$R_{\text{fm}} = R_{\text{self}}$
"**C**ombination"	free parameter	$R_{\text{fm}} = R_{\text{self}}$

models where β was treated as a free parameter, making the final meta-level reward a weighted average of R_{env} and R_{self}.

The families of models considered are described in Table 2. Further, models within each of these families differed from each other in one or more of the following features with binary feature values (v_f): (1) whether they generated pseudo-rewards for each meta-action ($v_f = 1$); (2) whether the priors for the feature weight parameters (θ_k) were initialized with the weights of the features of the participant's strategy as inferred by the method from Jain et al. (2022) ($v_f = 1$) or with standard normal distributions ($v_f = 0$); and (3) whether the absence of R_{env} was construed as $R_{\text{env}} = 0$ ($v_f = 1$) or simply as an absent learning-signal ($v_f = 0$). All plausible combinations of these features were considered, resulting in a total of 28 models across all families.

Model Fitting The parameters of the models were the learning rate α, the decay factor γ, the inverse temperature $\frac{1}{\tau}$, and the priors on the strategy feature weights θ_k. For models that used pseudo-rewards, an additional free parameter for the weight of the pseudo-reward was fit. Finally, for models in which the weight of the meta-level reward, β, was not fixed, β was an additional free parameter. All free parameters were fit individually to each participant's process-tracing data. This entailed using the TPE algorithm (Bergstra, Bardenet, Bengio, & Kégl, 2011) to minimize the negative log likelihood of clicks performed by the participant. To assess reproducibility, we fitted two sets of parameters independently for each participant for each model.

Model Comparison To compare the relevance and explanatory power of particular features of models in explaining the participants' data, we used family-level Bayesian model selection (Penny, Stephan, Mechelli, & Friston, 2004) to compare groups of models that are defined by certain characteristic features.

For each model or family of models, Bayesian model selection outputs (1) the estimate of the proportion of participants whose data is best explained by that model ($\hat{P}$), and (2) the exceedance probability (p_x), which is the probability that more participants are best explained by this particular model or family of models than any other.

To assess the reproducibility of our results, we repeated the model selection procedure on 8800 data sets we bootstrapped by sampling random combinations of the two sets of fits we obtained for each model (Wehrens, Putter, & Buydens, 2000).

Results

Here, we report on the results averaged over the 8800 repetitions of model selection. The 95% confidence intervals of all reported statistics have a width of less than 0.002 (or 0.2%, where these are reported as percentages.)

We first compared the family of models that learn only from explicit feedback with models that use self-evaluation to some extent. We found that, in the control condition, the models that solely rely on feedback best explained more participants ($\hat{P} = 0.70$, $p_x = 1.0$) than the models that used some self-evaluation ($\hat{P} = 0.30$, $p_x = 0.0$). Contrary to this finding, the models that learn from self-evaluation provided the best explanation for 82.6% of the participants ($\hat{P} = 0.83$, $p_x = 1.0$) in the scarcity condition.

Having determined the importance of self-evaluation to learning without feedback, we asked *how* self-evaluations are used for learning. For this, we compared four families of models: state-of-the-art models that use only explicit feedback (type "**V**anilla"), models that always learn exclusively from self-evaluation (type "**P**rioritize **S**elf"), models that use self-evaluation only in the absence of feedback (type "**P**rioritize **E**nv"), and models that always use a weighted combination of feedback and self-evaluation (type "**C**ombination"). See Table 2 for a technical description of the model families and the abbreviations of their names.

We found that, in the scarcity condition, the most common way participants used self-evaluation was as an addition to external feedback. Namely, the "C" models best explained 35.2% of the participants in this condition ($p_x = 0.763$). The second most common learning mechanism was to use self-evaluation only in the absence of feedback, with "PS" models best explaining 26.3% of the participants ($p_x = 0.175$). "PE" models best explained only 21.6% of participants in the scarcity condition ($p_x = 0.052$), while "V" models, which don't use self-evaluation, accounted for only 16.9% of participants ($p_x = 0.01$). In the control condition, on the other hand, models that learned only from feedback ("V" and "PE") best explained a majority of the participants, namely 69.0% ($p_x = 1.0$). The remainder of the participants were explained almost evenly by the families of models that used self-evaluation even in the presence of feedback: "C" models explained 15.7% of participants ($p_x = 0.0$) and "PS" models 15.3% ($p_x = 0.0$). Figure 3 visually depicts the differential use of self-evaluation signals within and between conditions.

According to these results, self-evaluation enabled participants to learn better strategies even in the absence of feedback. On trials with no feedback, 83.1% of the participants in the scarcity condition used self-evaluation in some way. By contrast, on trials with feedback, only 56.7% of these participants used self-evaluation. Compared to this, even fewer

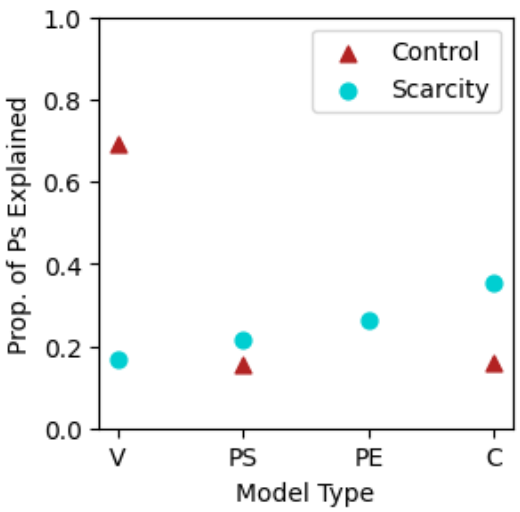

Figure 3: Proportions of participants in each condition best explained by different self-evaluation mechanisms. Only model type "V" does not use self-evaluations. When feedback is always present (Control condition), "PE" models are functionally similar to "V" models and are therefore included in the latter family. Refer to the text and Table 2 for descriptions of the model families and meanings of the abbreviations.

participants in the control condition used self-evaluation in the presence of feedback, namely only 29.7%.

Finally, we found that the other potential mechanism for learning in the absence of feedback – pseudo-rewards for gaining information – was not predominantly used by participants, though it was slightly more frequently used in the absence of feedback. The use of pseudo-rewards best explained the behavior of 40.2% ($p_x = 0.06$) of participants in the scarcity condition but only 24.8% ($p_x = 0.0$) in the control condition.

Discussion

How do people acquire cognitive skills? Previous work on metacognitive learning suggested simple reinforcement learning mechanisms (He et al., 2021). Our findings indicate that people rely on more sophisticated learning mechanisms, since the lack of feedback did not interfere with their ability to learn planning strategies through trial and error.

Consistent with previous findings on simpler forms of learning (Bandura, 1976), our model comparisons support the hypothesis that people boost their metacognitive learning by generating their own learning signals through self-evaluation (Andrade & Valtcheva, 2009). In our experiment, self-evaluation, though particularly important when feedback was absent, also contributed to strategy improvement even in the presence of feedback.

Overall, consistent with computational models of cognitive development (Rule, Tenenbaum, & Piantadosi, 2020; Shrager & Siegler, 1998), our results suggest that cognitive skill acquisition is more than simple reinforcement learning. Humans seem to learn more efficiently by giving themselves feedback on their own performance. In our experiment, they generated this feedback by estimating the expected utility of executing their plans from relevant information. These findings have implications for understanding and improving learning in humans as well as in computational agents. Firstly, we can use this knowledge to create learning environments that incentivize strategy improvement by facilitating self-evaluation, for example, by prompting people to engage in systematic reflection about their performance, their decisions, or their cognitive strategies (Wolfbauer, Pammer-Schindler, & Rosé, 2020; Becker, Wirzberger, Pammer-Schindler, Srinivas, & Lieder, 2023). Secondly, we could equip self-learning agents with methods for self-evaluation to make machine learning more sample-efficient and more robust in domains where feedback is scarce. Actor-critic reinforcement learning (Grondman, Busoniu, Lopes, & Babuska, 2012) could be seen as an example of that approach, but human self-evaluation is more sophisticated. Therefore, reverse-engineering people's capacity to boost their (metacognitive) learning through self-evaluation could be a promising way to advance artificial intelligence (Lake, Ullman, Tenenbaum, & Gershman, 2017).

The work presented in this article is only a first step toward understanding the computational mechanisms of how people discover adaptive cognitive strategies in feedback-scarce environments. Firstly, learning from self-evaluation in our experiment was straightforward. Whether self-evaluation plays a similarly prominent role in more naturalistic environments, where accurate self-evaluation is more challenging, remains to be seen. We expect that the importance of explicit feedback increases with the increase in the difficulty of assessing the quality of one's own plans (Kahneman & Klein, 2009). One such example is of doctors being uncertain about the effectiveness of their treatment plans when patients do not return or pass away in the short term (Omron, Kotwal, Garibaldi, & Newman-Toker, 2018).

Secondly, our models, which assume a single learning mechanism throughout the experiment, do not explain the additional finding that participants in the later stages of the experiment learned more from feedback trials than from non-feedback trials and learned more slowly in the feedback-scarce environment. This suggests that, while participants rapidly developed adaptive strategies through conceptual, insight-like learning in the first part of the experiment, they later relied on a more practice-based learning mechanism that is reliant on feedback (VanLehn, 1996).

Finally, we found that people are capable of a more sophisticated form of metacognitive learning than we had expected. Yet, a large proportion of participants didn't engage in it. This reinforces our hope that helping people tap into their capacity for deliberate metacognitive learning is a promising approach to improving their decision-making (Becker et al., 2023).

References

Anderson, J. R. (2013). *The adaptive character of thought*. Psychology Press.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice*, *48*, 12–19.

Bandura, A. (1976). Self-reinforcement: Theoretical and methodological considerations. *Behaviorism*, *4*(2), 135–155. Retrieved 2023-01-31, from http://www.jstor.org/stable/27758862

Becker, F., Wirzberger, M., Pammer-Schindler, V., Srinivas, S., & Lieder, F. (2023). Systematic metacognitive reflection helps people discover far-sighted decision strategies: A process-tracing experiment. *Judgment and Decision Making*.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Proceedings of the 24th international conference on neural information processing systems* (p. 2546–2554). Red Hook, NY, USA: Curran Associates Inc.

Brinko, T., Kathleen. (1993). The practice of giving feedback to improve teaching. *The Journal of Higher Education*, *64*, 574–593.

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.*

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

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.

Grondman, I., Busoniu, L., Lopes, G. A., & Babuska, R. (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 1291–1307.

He, R., Jain, Y. R., & Lieder, F. (2021). 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.*

He, R., & Lieder, F. (2022, jun). Where do adaptive planning strategies come from? Retrieved from https://doi.org/10.13140/RG.2.2.28966.60487 doi: 10.13140/RG.2.2.28966.60487

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*.

Jain, Y. R., Gupta, S., Rakesh, V., Dayan, P., Callaway, F., & Lieder, F. (2019). How do people learn how to plan? In *Conference on cognitive computational neuroscience (ccn 2019)* (pp. 826–829).

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American psychologist*, *64*(6), 515.

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

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

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*, e1.

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

Omron, R., Kotwal, S., Garibaldi, B. T., & Newman-Toker, D. E. (2018). The diagnostic performance feedback "calibration gap": Why clinical experience alone is not enough to prevent serious diagnostic errors. *AEM Education and Training*, *2*(4), 339–342.

Penny, W. D., Stephan, K. E., Mechelli, A., & Friston, K. J. (2004). Comparing dynamic causal models. *NeuroImage*, 1157–1172.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.

Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in cognitive sciences*, *24*(11), 900–915.

Shrager, J., & Siegler, R. S. (1998). Scads: A model of children's strategy choices and strategy discoveries. *Psychological science*, *9*(5), 405–410.

Siegler, R. S. (1999). Strategic development. *Trends in Cognitive Sciences*, *3*(11), 430–435.

Siegler, R. S., & Jenkins, E. A. (2014). *How children discover new strategies*. Psychology Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 643–662.

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology*, *47*, 513–539.

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

Wolfbauer, I., Pammer-Schindler, V., & Rosé, C. (2020). Rebo junior: analysis of dialogue structure quality for a reflection guidance chatbot. In *Ec-tel impact paper proceedings 2020: 15th european conference on technology enhanced learning.*