Measuring and modelling how people learn how to plan and how people adapt their planning strategies to the structure of the environment

Ruiqi He (ruiqi.he@tuebingen.mpg.de)
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Yash Raj Jain (yasshjain@gmail.com)
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Falk Lieder (falk.lieder@tuebingen.mpg.de)
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

Often we find ourselves in unknown situations where we have to make a decision based on reasoning upon experiences. However, it is still unclear how people choose which pieces of information to take into account to achieve well-informed decisions. Answering this question requires an understanding of human metacognitive learning, that is how do people learn how to think. In this study, we focus on a special kind of metacognitive learning, namely how people learn how to plan and how their mechanisms of metacognitive learning adapt the planning strategies to the structures of the environment. We first measured people's adaptation to different environments via a process-tracing paradigm that externalises planning. Then we introduced and fitted novel metacognitive reinforcement learning algorithms to model the underlying learning mechanisms, which enabled us insights into the learning behaviour. Model-based analysis suggested two sources of maladaptation: no learning and reluctance to explore new alternatives.

Keywords: decision-making; planning; metacognitive learning; reinforcement learning; cognitive modelling

Introduction

In real life, we often have to make decisions in new situations. Often our decisions and actions result from learned experiences and reasoning upon them. However, it is still unknown how we learn which pieces of information we should take into account to efficiently make a well-informed decision. Answering this question requires understanding how people learn how to think (metacognitive learning). While direct decision-making has been studied extensively from the perspective of cognitive science (Wang & Ruhe, 2007) and machine learning (Niv, 2009), our contemporary understanding of how people learn how to decide remains shallow. There is some work on modelling how people learn to select between the decision-making strategies they already know (Lieder & Griffiths, 2017; Rieskamp & Otto, 2006; Erev & Barron, 2005) but there is little work on how people discover those decision strategies in the first place. In this study, we focus on a special kind of metacognitive learning, namely how people learn how to plan.

Our work is structured in two parts - measuring and then modelling metacognitive learning in terms of reinforcement learning algorithms. For this, we set up an experiment that utilises a process-tracing paradigm that makes planning observable. The resulting process-tracing data is then analysed by a recently developed computational method for inferring people's planning strategies and their changes over time. To model how people learn how to plan, we formalised and tested three competing hypotheses about how people learn how to plan using three novel computational models. We tested our models against each other. The resulting best model was used to draw conclusions for different groups of participants.

By advancing our understanding of human metacognitive learning, this line of work may contribute to laying the foundations for improving metacognitive learning and helping people overcome maladaptive ways of decision-making.

Background

Mouselab MDP paradigm

A major obstacle to studying metacognitive learning is that we cannot directly observe people's cognitive strategies and how they change over time. To overcome this hurdle, we utilise a process-tracing paradigm that renders people's behaviour highly diagnostic of their planning strategies, namely the Mouselab Markov Decision Process (MDP) paradigm (Callaway, Lieder, Krueger, & Griffiths, 2017). In this paradigm, participants plan the route of a spider through a maze with the goal to maximise their score (see Figure 1) with the given number of trials. The score is the sum of the values of the nodes (the gray circle) on the path they choose to traverse. Each node harbours a gain or a loss, which are initially hidden but can be revealed by clicking on it. This explicit clicking action corresponds to evaluating the quality of a potential future state, which is a fundamental cognitive operation in planning. The cognitive cost of this operation is externalised by charging a fee of -1 for each node they reveal. Participants are thus encouraged to not immediately click every location, but instead, reveal information as necessary. In this way, the paradigm externalises the mental representation that people use for planning in terms of which nodes have been clicked and what their revealed values are.

Measuring metacognitive learning

The Mouselab-MDP paradigm can be used to measure the changes in people's strategy sequence. For this, we have previously developed a computational method that infers which planning strategy a participant used on each trial based on their clicks (Jain, Callaway, & Lieder, 2019; Jain et al., 2021). This method returns which of 79 predefined planning strategies a participant is most likely to have

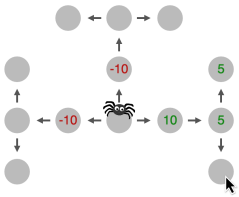


Figure 1: Example of the Mouselab paradigm for the constant variance condition with five nodes revealed

used. For detailed documentation of all 79 strategies please see `https://osf.io/zgshx/`. We can therefore measure metacognitive learning in terms of how the inferred strategy changed from each trial to the next.

Modelling metacognitive learning

To model metacognitive learning we will apply reinforcement learning algorithms to the problem of deciding how to decide (*meta-decision-making*). We will briefly introduce these two frameworks and how they can be combined.

Reinforcement learning Research suggests that human learning is partly driven by rewards and punishments, which they receive through trial and error (Niv, 2009). This learning mechanism has inspired reinforcement learning algorithms, which learn to estimate how much reward can be expected to receive from a certain action (a) in a given state (s). This estimate is updated according to the differences between received and predicted rewards δ:

$$Q(s,a) \leftarrow Q(s,a) - \alpha \cdot \delta \tag{1}$$

where α is the learning rate. To balance exploitation and exploration, the agent can choose its actions *probabilistically*, maximising the predicted action value, for example using the softmax rule (see for example Equation 3).

Meta-decision-making Previous work suggests that the brain is equipped with multiple decision systems that interact in numerous ways (Dolan & Dayan, 2013; Daw, 2018). In contrast to the Pavlovian and model-free systems, the model-based system supports flexible reasoning about which action might be best given available information, goals and preferences. To efficiently balance decision quality and decision time given enormous amount of information, the model-based system's flexibility necessitates a mechanism for selecting only relevant information, that is deciding how to decide, which is formally known as *meta-decision-making* (Boureau, Sokol-Hessner, & Daw, 2015). Recent work has formalised the problem of meta-decision-making as a meta-level MDP (Krueger, Lieder, & Griffiths, 2017; Griffiths et al., 2019):

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

where belief states $b_t \in \mathcal{B}$ encode the model-based decision system's beliefs about the values of alternative courses of actions. The temporal evolution of those belief states $(b_1, b_2, \cdots)$ is driven by the decision system's computations $c_1, c_2, \cdots$ according to the meta-level transition probabilities $T(b_t, c_t, c_{t+1})$. Finally, the meta-level reward function $r_{\text{meta}}(b_t, c_t)$ encodes the cost of performing the planning operation $c_t \in \mathcal{C}$ and the expected return of terminating planning $(c_t = \perp)$ and acting based on the current belief state b_t.

Metacognitive reinforcement learning Planning strategies can be thought of as policies for solving metalevel MDPs. We can therefore formalise the problem of discovering effective planning strategies as solving a metalevel MDP for the optimal metalevel policy (Griffiths et al., 2019). Solving meta-decision-making problems optimally is often computationally intractable but the optimal solution can be approximated through reinforcement learning (Russell & Wefald, 1991; Callaway, Gul, Krueger, Griffiths, & Lieder, 2018). Hence, we assume that the brain approximates optimal meta-decision-making via reinforcement learning mechanisms that seek to approximate the optimal solution of the meta-level MDP defined in Equation 2 by either learning to approximate the optimal policy directly or by learning an approximation to its value function. Previous work has applied this idea to model how people learn to select between alternative cognitive strategies (Erev & Barron, 2005; Rieskamp & Otto, 2006; Lieder & Griffiths, 2017), learn how many steps to plan ahead (Krueger et al., 2017), and learn when to exert how much cognitive control (Lieder, Shenhav, Musslick, & Griffiths, 2018). However, this approach has yet to be applied to investigate how people discover and refine their cognitive strategies.

Experiment

To investigate metacognitive learning, we designed an experiment with three conditions using the Mouselab-MDP paradigm to measure how people adapt their planning strategies to different environments.

Methods

Each participant was randomly allocated to one of three conditions. Each condition presented the participants with a different environment. In the increasing variance environment, the range of possible rewards is larger at locations that are further away from the starting point at the centre of the maze. In the decreasing variance environment, the variance between possible node values decreases the further away from the starting point, that is the nodes that are closest to the centre have the largest range of possible values. In the constant variance environment, the variance between possible node values remains the same. The possible value of each node at any given step can be seen in table 1. Step 1 corresponds to the three nodes that are closest to the starting point in the middle, step 2 is the next node, step 3 is the last set of nodes that are furthest away from the starting point.

Environment	Step 1	Step 2	Step 3
Increasing	-4, -2, 2, 4	-8, -4, 4, 8	-48, -24, 24, 48
Decreasing	-48, -24, 24, 48	-8, -4, 4, 8	-4, -2, 2, 4
Constant	-10, -5, 5, 10	-10, -5, 5, 10	-10, -5, 5, 10

Table 1: Possible reward values for the three environments

Participants We recruited 174 participants, 58 for each condition, on CloudResearch. The recruitment was limited to participants who had completed 100+ HITS, had a score > 90, and were located in the United States. Each participants received a base-pay of $1.50 and a bonus up to $5 based on their performance. They received minimal instructions and had to pass a quiz to demonstrate correct comprehension of the setup before starting the first trial.

Procedure Each participant was assigned to one condition was asked to complete 35 trials. The scores are displayed on the screen and are updated after each action (click, move). Planning is encouraged by a performance-depend bonus, which is 0.2 cents for each point of their final score after completion of all trials.

Results

To investigate whether people learn to adapt their planning strategies to the structure of the environment, the strategy sequences were analysed. To classify our participants' planning strategies into adaptive and maladaptive ones, we first created a list of planning strategies that were used by the participants and then determined the expected score of the strategies in the list using computer simulations. For each environment, the five strategies with the highest score are labelled as adaptive, while the five low scoring strategies are labelled as maladaptive strategies. We illustrated (see Figure 2) and tested the aggregated proportion of the five adaptive and five maladaptive strategies for trends using Mann Kendall tests. The tests confirmed an increasing trend for the aggregated proportion of adaptive strategies in all environments ($S > 367, p < .001$ in all environments). In addition, the tests suggest a decreasing trend for the maladaptive strategies in the increasing ($S = -429, p < .001$) and decreasing variance environments ($S = -295, p < .001$) and no trend in the constant variance environment ($S = -83, p = .176$). This means that in all three environments the proportion of people who adopted using adaptive strategies gradually increased while the proportion using maladaptive strategies gradually decreased in all but one environment. Furthermore, for each of those five adaptive and five maladaptive strategies, we tested whether the proportion of people using that strategy increased or decreased across trials using a series of Mann Kendall tests (see `https://osf.io/zgshx/` for detailed results of the tests). Overall, the tests suggested an increasing trend or no trend for the adaptive strategies, while the data indicated decreasing trend or no trend for the maladaptive strategies (see Figure 3). For instance, for the increasing variance condition, we found that the frequency of the adaptive strategy to search the final destinations for the best possible outcome (Strategy 21) steadily increased (Mann Kendall test: $S = 535, p < .001$), while the frequency of the maladaptive strategy to act without planning (Strategy 30) steadily decreased (Mann Kendall test: $S = -414, p < .001$).

These results suggest that people discover and learn to use adaptive strategies in all three environments. The effect is most prominent in the increasing variance condition and least prominent in the constant variance condition. This might be because discovering adaptive strategies is easiest when the environment has a clear structure that adaptive strategies can exploit.

Modelling metacognitive learning

Having empirically demonstrated that people discover and learn to use adaptive planning strategies, we now model the underlying computational mechanisms in terms of metacognitive reinforcement learning using two novel models: Learned value of computation (LVOC), direct adjustment of decision- making policy via gradient ascent (REINFORCE) and its non-learning variant, which postulates that people do not update their planning strategy. Each of these three models hypothesise a different learning mechanism.

Models of metacognitive reinforcement learning

Representations of the strategies The strategies people use in the Mouselab-MDP can be described in terms of a weighted combination of neuroscience-informed features. One example of a group of features are pruning features, which are related to assigning a negative value to thinking about a path whose expected value is below a certain threshold. Therefore, the learning trajectory can be expressed by how the weights of those features evolve over time. We have defined 52 different features (see `https://osf.io/zgshx/` for a detailed description).

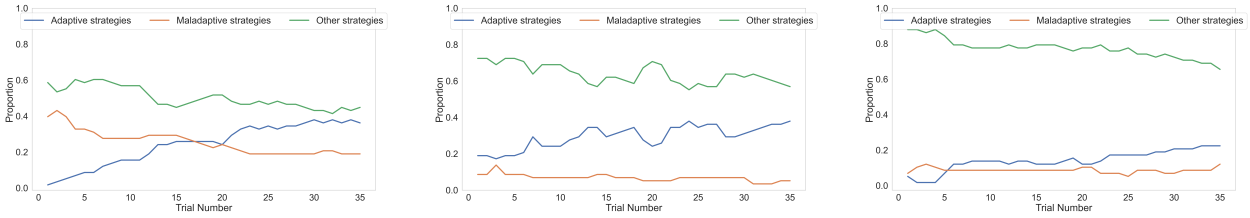
The REINFORCE model The REINFORCE model assumes that people adjust their planning strategy directly by following its performance gradient ascent through the strategy space using a softmax policy (Williams, 1992):

$$\pi_\theta(c|b) = \frac{\exp\left(\frac{1}{\tau} \cdot \sum_{k=1}^{52} \theta_k \cdot f_k(b,c)\right)}{\sum_{c \in C_b} \exp\left(\frac{1}{\tau} \cdot \sum_{k=1}^{52} \theta_k \cdot f_k(b,c)\right)} \quad (3)$$

where b is the belief state, c is the click being considered and C_b is the set of clicks available in the belief state b. τ is the inverse temperature parameter and f_k are the neuroscience-informed features values described above. The larger the value of τ is, the more deterministically the highest value action is chosen. The parameters of the policy (θ) are updated once after each trial according to the learning rule:

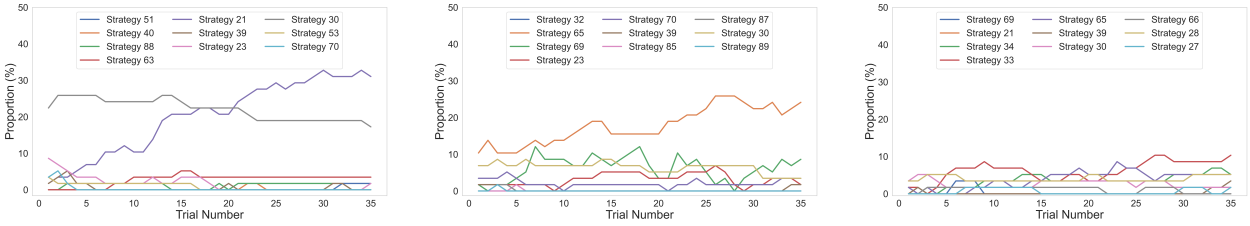
$$\theta \leftarrow \theta + \alpha \cdot \sum_{t=1}^{O} \gamma^{t-1} \cdot r_{\text{meta}}(b_t, c_t) \cdot \nabla_\theta \ln \pi_\theta(c_t|b_t) \quad (4)$$

where α is the learning rate, γ is the discount factor and O is the number of planning operations the model performed on



(a) Increasing variance environment (b) Decreasing variance environment (c) Constant variance environment

Figure 2: Proportion of aggregated strategy development throughout the trials for each environment.



(a) Increasing variance environment (b) Decreasing variance environment (c) Constant variance environment

Figure 3: Trial-wise proportion of adaptive and maladaptive strategy for each environment. The first 5 strategies are adaptive strategies, the last 5 strategies are maladaptive strategies

that trial, that is the number of clicks plus one, which represents the termination of planning operation. The learning rate α was optimised using ADAM (Kingma & Ba, 2014). Both α and γ are treated as free parameter that are fit separately for each participant. In addition to the vanilla REINFORCE, a pseudo-reward (Ng, Harada, & Russell, 1999) is used to speed up learning. The value of the pseudo-reward on performing computation c_t in belief state b_t and transitioning to belief state b_{t+1} is given by

$$\mathrm{PR}(b_t, c, b_{t+1}) = \mathbb{E}[R_{\pi_{b_{t+1}}} | b_{t+1}] - \mathbb{E}[R_{\pi_{b_t}} | b_{t+1}] \qquad (5)$$

which is the difference between the expected value of the best path in belief state b_{t+1} according to the policy $\pi_{b_{t+1}}$ and the expected value of the best path in belief state b_{t+1} according to the policy π_{b_t}.

The LVOC model According to the LVOC model, people discover and change their strategy continuously by learning to predict the values of alternative planning operations (Krueger et al., 2017). The model assumes that people learn a linear approximation to the meta-level Q-function:

$$Q_{\mathrm{meta}}(b_t, c_t) \approx \sum_{k=1}^{52} w_k \cdot f_k(b_t, c_t), \qquad (6)$$

using the f_k and corresponding weights w_k The LVOC model learns the weights w_k of those features by Bayesian linear regression of the bootstrap estimate $\hat{Q}(b_t, c_t)$ of the meta-level value function onto the features f_k. The bootstrap estimate

$$\hat{Q}(b_t, c_t) = r_{\mathrm{meta}}(b_t, c_t) + \sum_{k=1}^{52} \mu_{k,h} \cdot f_k(b_{t+1}, c_{t+1}) \qquad (7)$$

is the sum of the immediate meta-level reward and the predicted value of the next belief state b_{t+1} under the current meta-level policy. The predicted value of b_{t+1} is the scalar product of the posterior mean $\mu_{k,h}$ of the weights w_k, given the observations from all h preceding planning operations and the features $f_k(b_{t+1}, c_{t+1})$ of b_{t+1} and the cognitive operation c_{t+1} that the current policy selects given state. Given the posterior on the feature weights $\mathbf{w} = (w_1, ..., w_{52})$, the next planning operation c_{t+1} is selected by a generalised version of Thompson sampling. That means, to make the k^{th} meta-decision, n weight vectors $\tilde{w}^{(1)}, \cdots, \tilde{w}^{(n)}$ are sampled from the posterior distribution of the weights given the series of meta-level states, selected planning operations, and resulting value estimates experienced so far. That is,

$$\tilde{w}_t^{(1)}, \cdots, \tilde{w}_t^{(n)} \sim P(\mathbf{w} | \mathcal{E}_t), \qquad (8)$$

where the set $\mathcal{E}_k = \{e_1, \cdots, e_t\}$ contains the meta-decision-maker's experience from the first t meta-decisions. To be precise, each meta-level experience $e_j \in \mathcal{E}_k$ is a tuple $\left(b_j, h_j, \hat{Q}(b_j, c_j; \mu_j)\right)$ containing a meta-level state, the computation selected in it, and the bootstrap estimates of its Q-value. The arithmetic mean of the sampled weight vectors $\tilde{w}^{(1)}, \cdots, \tilde{w}^{(n)}$ is then used to predict the Q-values of each possible planning operation $c \in \mathcal{C}$ according to Equation 6. The planning operation with the highest predicted Q-value is used for decision-making. For a fair comparison, the LVOC model also utilises the metacognitive pseudo rewards defined in Equation 5. The LVOC model has three free parameters: the mean vector $\boldsymbol{\mu}_{prior}$ and variance $\sigma^2_{\mathrm{prior}}$ of its prior $\mathcal{N}(\mathbf{w}; \boldsymbol{\mu}_{prior}, \sigma^2 \cdot \mathrm{I})$ on the weights $\mathbf{w}$ and the number of samples n.

Model fitting methods

To assess how well each model can capture how people learn how to plan, we fitted each model's free parameters and priors on feature weights to the participant's data and applied each model to the series of problems the participant had to solve.

The parameters of the models were fit by maximising a Multivariate-Normal pseudo-likelihood function defined in terms of the probability that the model would generate the participant's trial wise scores as a function of its parameters. For a given participant i, the pseudo-likelihood function under model m is given by:

$$\mathcal{L}((\theta_{i,m}, \sigma_{i,m})|\mathbf{r_i}) = \phi(\mathbf{r_i}; \hat{\mathbf{r}}_{\mathbf{i,m}}(\theta), \sigma_{i,m} I) \tag{9}$$

where $\theta_{i,m}$ is the parameter vector used to fit the data from participant i with model m, $\mathbf{r_i}$ is the vector of scores that the i^{th} participant obtained on trials 1 through 35, σ is the standard deviation of the errors between the observed scores and the model's predictions $\hat{\mathbf{r}}_{\mathbf{i,m}}(\theta_{i,m})$, and $\phi(\mathbf{x}; \mu, \Sigma)$ is the density function of the multivariate normal distribution. We estimate the parameters $\theta_{i,m}$ and $\sigma_{i,m}$ by maximising the pseudo-likelihood function in Equation 9 using Bayesian Optimisation (Bergstra, Yamins, & Cox, 2013). All selected models are then fit to the participant data using 400 iterations. In each iteration, the model's prediction is estimated by averaging the model's scores across 30 simulations.

Model selection

After the model-fitting, we performed individual-level and group-level model selection using the Akaike Information Criterion (AIC) (Akaike, 1998). On the level of individual participants, both learning models, LVOC and REINFORCE, seem to explain the learning behaviour reasonably better than the non-learning model (see Table 2). The number of participants whose data was best explained was the same for both learning models (71). To investigate the differences in

Environment	Model	Count
Increasing variance	non-learning	11
	REINFORCE	24
	LVOC	23
Decreasing variance	non-learning	10
	REINFORCE	28
	LVOC	20
Constant variance	non-learning	11
	REINFORCE	19
	LVOC	28

Table 2: Count of individual participants' best fitted model.

which model explains a participant's data best, we divided the participants into three groups: participants who were not using adaptive strategies in the beginning but learned to do so were classified as *highly adaptive learners*, participants using maladaptive strategies in the last trial were classified as *maladaptive participants*, and the other participants are labelled as *moderately adaptive participants*. The group-level model comparison provided strong evidence in favour of the REINFORCE model (average AIC = 308.31) over the LVOC model (average AIC = 315.94) and over its non-learning variant (average AIC = 341.43). As shown in Figure 4, the REINFORCE model was able to capture how the participants' performance throughout the experiment in all three conditions. Most importantly, the REINFORCE model was able to capture the improvement in people's performance as they adapt their planning strategies to the structure of the increasing variance environment (Figure 4a).

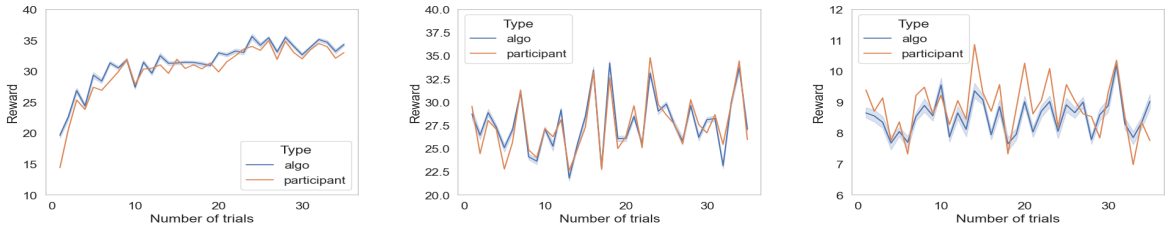
Increasing (n=58)	non-learning	REINFORCE	LVOC
Highly adaptive (n=21)	387.44	**343.54**	346.97
Maladaptive (n=11)	184.42	**174.68**	205.36
Mod. adaptive(n=26)	375.56	**341.55**	351.25
Decreasing (n=58)	non-learning	REINFORCE	LVOC
Highly adaptive (n=16)	369.84	326.66	**320.53**
Maladaptive (n=3)	202.95	198.94	**197.86**
Mod. adaptive (n=39)	370.27	**306.39**	324.88
Constant (n=58)	non-learning	REINFORCE	LVOC
Highly adaptive (n=11)	349.30	**330.33**	334.64
Maladaptive (n=7)	326.66	316.72	**309.72**
Mod. adaptive (n=40)	307.08	**290.15**	294.28

Table 3: Averaged AIC for each model grouped by participants. Best performance is marked in bold.

Model-based analysis

Due to its superior performance, REINFORCE was chosen to perform model-based analysis to gain insights into the learning behaviour and more specifically how they differ among groups of participants.

We hypothesised that maladaptive participants would have lower learning rates than the other two groups and tested our hypothesis using Wilcoxon rank-sum tests on the fitted learning rates. In addition, exploratory Wilcoxon rank-sum tests were conducted on the other parameters γ, which quantifies the influence of immediate meta-level rewards as opposed to the reward received later during the trial, and τ, which describes to which extent the participant explores different strategies (see Table 4). For the increasing variance environment, the tests imply that the distribution of inverse temperature parameters differs significantly between maladaptive ($M = 233.94, SD = 380.68$) and moderately adaptive participants ($M = 37.89, SD = 88.61$). This suggests that maladaptive participants might choose their planning operations more deterministic and thereby perform less cognitive exploration of alternative planning strategies. Participant-level analyses confirmed that 9 out of the 11 maladaptive participants started with a maladaptive strategy and either did not change their strategy or only changed it once. This suggests that the reason why some people find it difficult to steer away from maladaptive decision strategies is that they fail to explore alternative decision strategies. In the decreasing variance environment, the learning rate also differed significantly between maladaptive participants ($M < 0.0001, SD < 0.0001$) and the other two groups (highly adaptive: $M = 0.007, SD = 0.018$; moderately adaptive: $M = 0.009, SD = 0.026$). The small learning rate suggests that maladaptive participants did not learn at all. In the constant variance environment, the significant



(a) Increasing variance environment (b) Decreasing variance environment (c) Constant variance environment

Figure 4: Averaged REINFORCE model performance visualised against participants' performance

difference in inverse temperature implies that highly adaptive learners ($M = 9.12, SD = 26.51$) explore more than the maladaptive ones ($M = 32.46, SD = 54$), which aligns with the adaptive strategy for this environment.

Parameter	Comparison	T	p
Inverse temperature (increasing variance)	Malad. vs. Mod. ad.	2.79	.005
Inverse temperature (constant variance)	Malad. vs. Highly. ad.	2.22	.026
Learning rate (decreasing variance)	Malad. vs. Highly. ad.	-2.01	.044
	Malad. vs. Mod. ad.	-1.98	.048

Table 4: Results of Wilcoxon rank sum test on the fitted parameters

Conclusion and further work

We first measured how people adapt their planning strategies to different environments and then modelled the underlying learning mechanisms in terms of metacognitive reinforcement learning. Using a process-tracing method, we found that participants discovered different types of planning strategies depending on what was adaptive for the environment they were in. Concretely, the proportion of adaptive strategies significantly increased in all environments, while the proportion of maladaptive strategies significantly decreased in almost all environments. After having confirmed that people adapt to all three environments, we proceeded to develop and test two new models of metacognitive reinforcement learning. Our models extend previous models of metacognitive learning (Lieder & Griffiths, 2017; Krueger et al., 2017; Lieder et al., 2018) to the problem of strategy discovery. They achieve this by learning a policy for selecting individual planning operations. In additional, innovation of our models is that they learn not only from external rewards but also from intrinsically generated pseudo-rewards for gaining valuable information. Model selection suggested that the REINFORCE model best describes how people learn how to plan. Our model-based analysis of individual differences in metacognitive learning highlighted two potential causes of maladaptative planning: no learning and reluctance to explore. Further work could look into how to motivate learning and exploration - for example by gamification.

Acknowledgement

We thank Valkyrie Felso for code-related help and Frederic Becker for his support in demonstrating how to run experiments.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer.

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning* (pp. 115–123).

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.

Callaway, F., Gul, S., Krueger, P., Griffiths, T. L., & Lieder, F. (2018). Learning to select computations. In *Uncertainty in artificial intelligence: Proceedings of the thirty-fourth conference.*

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.*

Daw, N. D. (2018). Are we of two minds? *Nature Neuroscience*, *21*(11), 1497.

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

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, *112*(4), 912.

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.

Jain, Y. R., Callaway, F., Griffiths, T. L., Dayan, P., Krueger, P. M., & Lieder, F. (2021). A computational process-tracing method for measuring people's planning strategies and how they change over time.

Jain, Y. R., Callaway, F., & Lieder, F. (2019). Measuring how people learn how to plan. In *Cogsci* (pp. 1956–1962).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

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

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

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml* (Vol. 99, pp. 278–287).

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Rieskamp, J., & Otto, P. E. (2006). Ssl: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207.

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

Wang, Y., & Ruhe, G. (2007). The cognitive process of decision making. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *1*(2), 73–85.

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