A resource-rational analysis of human planning

Frederick Callaway[1], Falk Lieder[1], Priyam Das, Sayan Gul, Paul M. Krueger & Thomas L. Griffiths

Department of Psychology, University of California, Berkeley

[1] These authors contributed equally.

Abstract

People's cognitive strategies are jointly shaped by function and computational constraints. Resource-rational analysis leverages these constraints to derive rational models of people's cognitive strategies from the assumption that people make rational use of limited cognitive resources. We present a resource-rational analysis of planning and evaluate its predictions in a newly developed process tracing paradigm. In Experiment 1, we find that a resource-rational planning strategy predicts the process by which people plan more accurately than previous models of planning. Furthermore, in Experiment 2, we find that it also captures how people's planning strategies adapt to the structure of the environment. In addition, our approach allows us to quantify for the first time how close people's planning strategies are to being resource-rational and to characterize in which ways they conform to and deviate from optimal planning.

Keywords: bounded rationality; planning; rational analysis; decision-making; heuristics

Introduction

Previous research has shown that many aspects of human cognition can be understood as rational adaptations to the environment and the goals people pursue in it (Anderson, 1990). *Rational analysis* leverages this assumption to derive models of human behavior from the structure of the environment. In doing so, rational analysis makes only minimal assumptions about cognitive constraints. However, it has been argued that there are many cases where the constraints imposed by cognitive limitations are substantial, and Herbert Simon famously argued that to understand people's cognitive strategies we have to consider both the structure of the environment and cognitive constraints simultaneously (Simon, 1956, 1982). *Resource-rational analysis* (Griffiths, Lieder, & Goodman, 2015) thus extends rational analysis to also take into account which cognitive operations are available to people, how long they take, and how costly they are. Given that resource-rational analysis has been successful at explaining a wide range of cognitive biases in judgment (Lieder, Griffiths, Huys, & Goodman, 2017) and decision-making (Lieder, Griffiths, & Hsu, 2018) by suggesting resource-efficient cognitive mechanisms, it might also be able to shed new light on other cognitive processes, such as planning.

Surprisingly little is known about how people plan. While extant models of planning (De Groot, 1965; Huys et al., 2012, 2015; Newell & Simon, 1956, 1972) explain aspects of human planning, its precise mechanisms remain unclear; the applicability of each existing model is limited; and it remains unknown when people use which of those strategies and why. These questions are very difficult to answer because planning is an unobservable and highly complex cognitive process.

Here, we address these problems by deriving planning strategies through resource-rational analysis and introducing a process-tracing paradigm that allows us to directly observe the sequence of people's planning operations. We use data obtained with this paradigm to evaluate our resource-rational model of planning against previously proposed planning strategies. This enables us to distinguish between those models even when they predict the same final decision.

Our resource-rational framework enables us to automatically discover the optimal planning strategy for any given environment. We find that people's planning strategies are better explained by bounded-optimal planning than by classic models of planning as search (progressive deepening, best-first search, depth-first search, and breadth-first search) even when those models are augmented with the mechanisms of satisficing (Simon, 1956) and pruning (Huys et al., 2012). We characterize how human planning conforms to and deviates from resource-rational planning and quantify individual differences in the rationality of people's planning strategies. Furthermore, our analysis correctly predicts how people's planning strategies differ across environments.

This paper is structured as follows. We start by introducing the methodology of resource-rational analysis and review previous findings on planning. Next, we introduce our new process-tracing paradigm for the study of planning and apply resource-rational analysis to its planning problems. We then evaluate the resource-rational model against process-tracing data from people in Experiment 1. Experiment 2 tests resource-rational predictions about how people's planning strategies should change with the structure of the environment. We close by discussing the implications of our findings for cognitive modeling and human rationality.

Background

Discovering optimal cognitive strategies

Resource-rational analysis (Griffiths et al., 2015) derives process models of how cognitive abilities are realized from a formal specification of their function and a model of the cognitive architecture available to realize them. Formally, the resource-rational model of a cognitive mechanism is defined as the solution to a constrained optimization problem over the space of strategies that can be implemented on the assumed cognitive architecture, and the objective function measures how well the strategy would perform under the constraints of limited time and costly computation. This problem formulation can be approximated as a meta-level Markov decision process (Hay, Russell, Tolpin, & Shimony, 2012).

Planning

Most research on planning has been conducted in the fields of problem solving and artificial intelligence (Newell & Simon, 1972). The Logic Theorist (Newell & Simon, 1956) planned its proofs using *breadth-first search*: it first evaluated all possible one-step plans, then proceeded to all possible two-step plans, and so on, until it discovered a proof. By contrast, chess programs typically use *depth-first search*: they evaluate one possible continuation in depth and then back up one step at a time. When an optimal solution is not necessary (or feasible), an inadmissible heuristic can be applied: For example, (greedy) *best-first search* searches in whatever direction looks most promising at the moment.

Newell and Simon's (1972) research on human problem solving found that people usually plan forwards by a strategy called *progressive deepening* (De Groot, 1965) which is similar to depth-first search but resumes planning from the beginning after having considered one action sequence in depth. Furthermore, Simon (1956) argued that human decision-making is fundamentally constrained by limited cognitive resources and that people cope with these constraints by choosing the first option they find good enough instead of trying to find the best option; this is known as *satisficing*.

More recent work has found that people often prune their decision tree when they encounter a large loss (Huys et al., 2012) and cache and reuse previous action sequences (Huys et al., 2015). It has also been argued that people greedily choose each of their planning operations so as to maximize the immediate improvement in decision quality instead of considering the potential benefits of sequences of planning operations (Gabaix, Laibson, Moloche, & Weinberg, 2006).

The Mouselab-MDP paradigm

Planning, like all cognitive processes, cannot be observed directly. In previous work, researchers have inferred properties of human planning from the decisions participants ultimately made or asked participants to verbalize their planning process. However, many different planning strategies can lead to the same final decision, and introspective reports can be incomplete or inaccurate.

To address these challenges we employ a new *process-tracing paradigm* for the study of planning that externalizes people's unobservable beliefs and planning operations as observable states and actions (Callaway, Lieder, Krueger, & Griffiths, 2017). Inspired by the Mouselab paradigm (Payne, Bettman, & Johnson, 1993) that traces how people choose between multiple risky gambles, the Mouselab-MDP paradigm uses people's mouse-clicking as a window into their planning.

Each trial presents a route planning problem where each location (the gray circles in Figure 1), harbors a reward or punishment. These potential gains and losses are initially occluded, corresponding to a highly uncertain belief state, but the participant can reveal each location's value by clicking on it and paying a fee. This problem is equivalent to looking at a map and selecting a sequence of destinations for a road trip.

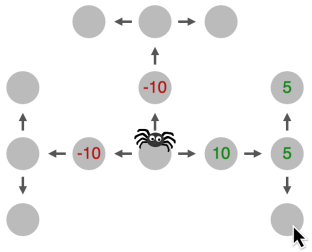

Figure 1: Illustration of the Mouselab-MDP paradigm. Rewards are revealed by clicking, prior to selecting a path with the arrow keys.

Clicking on a location roughly corresponds to thinking about a potential destination, evaluating how enjoyable it would be to go there, and adjusting one's assessment of trips including this destination accordingly. Although this is a dramatic simplification of the representations and computations people employ when planning, it nevertheless retains enough of the core structure of planning to reveal previously unobservable aspects of human planning.

Models of planning

We model planning in the Mouselab-MDP paradigm as a metalevel Markov Decision Process (metalevel MDP; Hay et al., 2012),

$$M_{\text{meta}} = (\mathcal{B}, \mathcal{A}, \mathcal{T}, r_{\text{meta}}), \tag{1}$$

where each belief state $b \in \mathcal{B}$ encodes the joint distribution over the rewards at each node (i.e. location) in the planning graph. In the case of uniformly distributed rewards, the belief state $b^{(t)}$ at time t can be represented as $\left(\mathcal{R}_1^{(t)}, \cdots, \mathcal{R}_K^{(t)}\right)$ where $\mathcal{R}_1^{(t)}$ is the set of possible values that the hidden reward X_k might take such that $b^{(t)}(X_k = x) = \text{Uniform}(x; \mathcal{R}_k^{(t)})$. The metalevel actions are $\mathcal{A} = \{c_1, \cdots, c_K, \perp\}$ where c_k reveals the reward at node k and $\perp$ selects the path with highest expected sum of rewards according to the current belief state. The transition probabilities $T_{\text{meta}}(b^{(t)}, c_k, b^{(t+1)})$ encode that performing computation c_k sets $\mathcal{R}_k^{(t+1)}$ to $\{x\}$ with probability $1/\left|\mathcal{R}_k^{(t)}\right|$ for $x \in \mathcal{R}_k^{(t)}$. The metalevel reward function is $r_{\text{meta}}(b, c) = -\lambda$ for $c \in \{c_1, \cdots, c_K\}$, and

$$r_{\text{meta}}\left((\mathcal{R}_1, \cdots, \mathcal{R}_K), \perp\right) = \max_{\mathbf{t} \in \mathcal{T}} \sum_{k \in \mathbf{t}} \frac{1}{|\mathcal{R}_k|} \cdot \sum_{x \in \mathcal{R}_k} x, \tag{2}$$

where $\mathcal{T}$ is the set of all paths $\mathbf{t}$.

Models

We model people's planning operations c as arising from a combination of a systematic strategy M and unexplained variability according to

$$\Pr(c|b, M, \varepsilon, \tau) = (1 - \varepsilon) \cdot \sigma(c; V_{b,M}, \tau) + \varepsilon \cdot \mathcal{U}(c; C_b), \tag{3}$$

where the first term models the strategy's choice of computations as a soft-max function ($\sigma(c; V_{b,M})$), and the second term

models unexplained variability by a uniform distribution over the set C_b of all clicks that have not been made yet and the termination action $\perp$. The weight ε of the random process is a free parameter that we constrain to be less than 0.25. The probability $\sigma(c; V_{b,M})$ that the strategy will choose computation c is defined as a soft-max decision rule

$$\sigma(c; V_{b,M}, \tau) = \frac{\exp(\frac{1}{\tau} \cdot V_{b,M}(c))}{\sum_{c' \in C_b} \exp(\frac{1}{\tau} \cdot V_{b,M}(c'))}, \quad (4)$$

over its belief-dependent preferences $V_{b,M}$. The decision temperature τ interpolates between always choosing the most preferred computation and choosing computations at random. The models presented below differ only in $V_{b,M}(c)$.

Optimal planning The optimal planning strategy is the one that always takes the metalevel action with maximal expected long term reward. In standard MDP notation (Sutton & Barto, 1998), the expected long term reward of executing action a in state s (and then continuing optimally) is given by $Q^*(s,a)$; in a metalevel MDP, we simply replace actions a and states s with beliefs b and computations c. Thus, our model of optimal planning has $V_{b,\text{OPT}}(c) = Q^*(b,c)$. We compute Q^* exactly by backwards induction.

Classical planning strategies To evaluate our optimal planning strategy against extant theories, we built likelihood models of the classical planning strategies known as depth-first search, breadth-first search, best-first search, and progressive deepening search (Newell & Simon, 1972). We augment these classic search-based strategies with satisficing (Simon, 1956) and pruning (Huys et al., 2012) that allow the strategy to terminate planning before clicking every node. Thus, the preference function for each heuristic strategy is defined piecewise by one function that determines the search order (preferences for clicks, $V_b(c) \forall c \in C_b \setminus \{\perp\}$), and another that determines the termination criterion (preference for terminating, $V_b(\perp)$).

Beginning with the second piece, we assume that the heuristic strategies terminate search by a combination of satisficing and pruning. When the expected reward for terminating search given the current belief equals or exceeds the model's aspiration level α, then $V_{b,M}(\perp) = 10^{10}$ so that all strategies strongly prefer to terminate planning; otherwise, $V_{b,M}(\perp) = -10^{10}$ making termination undesirable. Terminating is still possible in this case by pruning: If the expected return of a path falls below the pruning threshold ω, then $V_{b,M}(c) = -10^{20}$ for all computations c that inspect any of the nodes along that path, making clicking much less desirable than terminating even when the aspiration level has not been met. The aspiration level α and the pruning threshold ω are free parameters, constrained to be strictly positive and strictly negative respectively (to exclude the degenerate case of always preferring to terminate search).

Turning now to the search order, we define $V_{b,M}$ for each model M such that $\sigma(c; V_{b,M}, \tau = 10^{-10})$ reproduces the behavior of the modeled strategy M. For example, the preference function for depth-first search has

$$V_{b,\text{DFS}}(c) = \begin{cases} -10^{10} & \text{if } \neg\text{clicked}(\text{parent}(c)) \\ \text{depth}(c) & \text{otherwise} \end{cases}, \quad (5)$$

where the first case makes the strategy search in traversal order and the second makes it prefer to click nodes that are further from the root of the tree. Breadth-first search prefers less distant nodes by simply negating $\text{depth}(c)$ in Equation 5. Best-first search prioritizes nodes on promising paths by replacing $\text{depth}(c)$ with the expected sum of rewards along the path on which c lies.

Directed cognition Finally, we considered an extension of the directed cognition model of Gabaix et al. (2006). The directed cognition model uses macro-operators, which we define as sequences of clicks along a path. Therefore, each macro-operator can be defined by a (p, n_c)-tuple, where p is a path and n_c is the number of clicks the macro-operator makes along that path. The nodes on the path are clicked in the order of decreasing reward variance, with ties broken at random. The directed cognition model chooses macro-operators according to a myopic cost-benefit analysis. Concretely, the value $V_{b,\text{DC}}(o)$ of a macro-operator o is the expected utility of terminating search immediately after executing o minus the utility of terminating immediately and the cost of executing o. Macro-operators are selected with noisy maximization as in Equation 3.

Experiment 1: Testing models of planning

In Experiment 1, we leveraged the Mouselab-MDP paradigm (Figure 1) to evaluate how people plan against optimal planning and classic models of planning.

Methods

Stimuli and Procedure The experiment began with a series of practice blocks that introduced the problem (one for navigating with all rewards revealed, one for navigating with all rewards concealed, one for inspecting nodes, and one that introduced the cost of inspecting rewards). Participants then took a quiz that queried participants about the range of rewards, the cost per click, and how their bonus would be calculated.

In the main part of the experiment each participant solved 30 different 3-step planning problems of the form shown in Figure 1. There were 3 options for the first move and two options for the third move, leading to 6 paths in total. Each location's reward was independently drawn from a discrete uniform distribution over the values $\{-10, -5, +5, +10\}$, and the cost of inspecting a node was $\lambda = 1$. This cost was deducted directly from the participant's earnings, or score, which was initialized at 50. To reduce the opportunity cost of time, participants were required to spend at least 7 seconds on each

Model	BIC	LL	R^2
Optimal	30625	-15303	0.115
Best First	31744	-15854	0.083
Breadth First	32387	-16176	0.064
Depth First	32454	-16209	0.062
Progressive Deepening	32476	-16220	0.062
Directed Cognition	34025	-17004	0.017
Random	34579	-17289	0

Table 1: Model comparison: Columns are Bayesian Information Criterion, Log Likelihood, and McFadden's pseudo R^2.

trial; if they finished the trial in less time, a countdown appeared and they were told to wait until the remaining time had passed.

Participants We recruited 60 participants from Amazon Mechanical Turk. Each participant received a base pay of \$0.50 and a performance-dependent bonus that was proportional to their score in the task (average bonus: $\$2.16 \pm \$1.16$) for about 16.6 minutes of work on average. We excluded 9 participants (15%) because they either failed to follow the instruction to click during the training phase or answered more than 1 of the 3 comprehension checks incorrectly.

Results

Overall, participants' average score of 6.54 ± 0.31 points per trial was about 70% of the average score achieved by the optimal planning strategy (9.33). The moves participants selected were optimal relative to the information they had uncovered during planning in 98.6% of the trials. The subsequent analyses therefore focus on people's planning strategy.

Model Comparisons We began by evaluating how well each model explains the aggregate data pooled across all participants. We fit each model's free parameters by maximum likelihood estimation. To account for the differing number of parameters, we computed the Bayesian Information Criterion (Schwarz, 1978). As shown in Table 1, our data provided strong evidence in favor of the optimal model in terms of the complexity-penalized BIC and the raw likelihoods.

Modeling process-tracing data at the resolution of individual planning operations is intrinsically difficult. To get a sense of the predictive power of each model over the random baseline, we compute McFadden's pseudo R^2 using the random model as the null model: $R^2(M) = 1 - \frac{LL(M)}{LL(M_{\text{random}})}$. We find that all models explain a relatively modest proportion of the variance. This is at least in part due to inherent limits to the predictability of our data, including the symmetry in the problem that often makes multiple clicks functionally equivalent and the individual differences documented below.

Next, we characterize in which ways people's strategies deviated from and conformed to optimal planning.

Qualitative predictions Inspecting the click sequences of the resource-rational strategy across 40 simulated trials suggested that it always inspects a node on a path with the highest expected return (of all paths with unclicked nodes)–like best-first search. People follow this pattern 82.1% of the time ($p < 10^{-15}$). However, unlike best-first search, the optimal strategy terminates search when the expected value of information drops below the cost of attaining it. As a result, the optimal aspiration level decreases as more information is acquired. As shown in Figure 2a, holding the value of the best path constant, the probability that the optimal strategy stops planning increases with the number of clicks already made. This pattern is noisily expressed in the human data as well: A mixed effects logistic regression of the termination probability on the number of revealed states and the value of the current best path revealed a significant negative interaction ($\chi^2(1) = 43.319, p < 10^{-10}$).

However, we also found two systematic deviations of human planning from resource-rationality: First, unlike the optimal strategy, people preferred to inspect states in the order they would traverse them. Concretely, people preferred inspecting the rewards for the first step over the rewards for the second step (72.6% vs. 10.0%, $p < .0001$) even though optimal planning is indifferent between them. Similarly, after observing a large positive reward for taking a certain action in the first step in their first click, people more often conformed to best-first search which evaluates the immediate next step (51.5%) than to the optimal strategy which skips ahead to one of its final destinations (28.9% of the time; $\chi^2(1) = 10.1, p < .0015$). These deviations might reflect that simulating actions in future states is more costly than simulating acting in the present. Second, following a moderately good observation on the first click, people continue to explore paths starting with other actions 57.0% of the time whereas the resource-rational strategy would zoom in on the most promising paths identified by that observation.

Quantifying deviations from bounded optimality. We found that, on average, 45.0% of our participant's computations were sub-optimal. However, the computations people selected did nevertheless achieve 86.8% of the highest possible value of computation ($\text{VOC}(b,c) = Q^\star_m(b,c) - Q^\star_m(b,\perp)$); we will refer to this ratio as the *rationality quotient*. Next, we characterized the ways in which people's planning strategies are sub-optimal. We found that people tend to plan too little. Concretely, 28.3% of people's deviations from optimal planning were caused by stopping too early, but only 6.3% were caused by stopping too late. Finally, the majority of people's deviations from bounded optimality (i.e., 65.5%) occurred when they clicked on one node when the optimal strategy would have clicked on a different node.

Individual differences in rationality. Consistent with previous work by Stanovich and West (1998) we found considerable inter-individual differences in the extent to which peo-

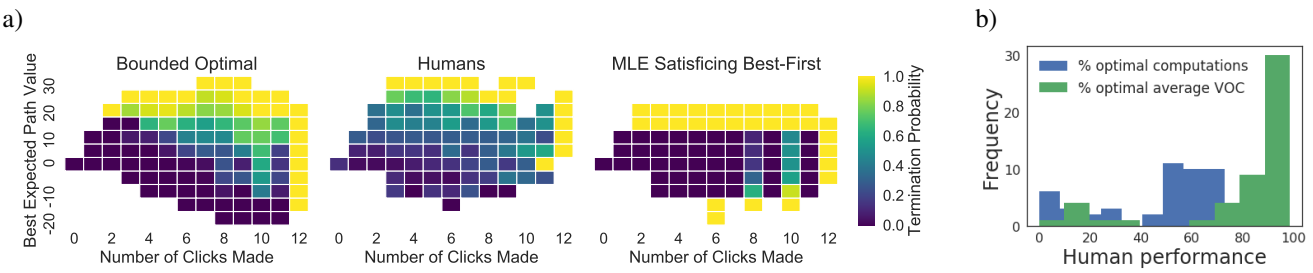

Figure 2: a) Adaptive termination threshold in optimal and human planning. Both the optimal model and humans smoothly lower the termination threshold as more clicks are made. The next-best-fitting model does not have this characteristic. b) Individual differences in the rationality of people's planning strategies.

ple's planning strategies were rational (see Figure 2b). The average agreement between people's planning operations and optimal planning ranged from 0% to 80.8% (48.0% ± 23.3%); and the average VOC of people's planning operations ranged from 0% to 98.5% of the optimal VOC (80.0% ± 26.4%). Figure 2b suggests that the majority of participants achieved a high level of resource-rationality; very high rationality quotients were more common than low ones, and the distribution of people's rationality scores might be bimodal.

Experiment 2: Structure shapes strategies

Bounded optimality predicts that people should adapt their planning strategy to the structure of the environment. We test this prediction by manipulating whether future rewards are more variable than immediate rewards or vice versa. Because high-variance rewards have a greater effect on a path's total value, the optimal strategy will inspect these nodes first. Thus, it will plan forwards when reward variance decreases with depth and backwards when reward variance increases with depth. In Experiment 2, we test whether human planners are likewise sensitive to the structure of reward variance.

Methods

Experiment 2 presented participants with a modified version of the three-step planning task from Experiment 1 (see Figure 1). Each participant was randomly assigned to one of two conditions that differed in whether the variability of a node's reward distribution increases or decreases with the number of steps it takes to reach that node. Concretely, in the first condition the reward distributions were Uniform($\{-4,-2,+2,+4\}$), Uniform($\{-8,-4,+4,+8\}$), and Uniform($\{-48,-24,+24,+48\}$) for nodes reachable in one, two, and three steps, respectively. In the second condition, the order of these distributions was reversed. The instructions informed participants about this reward structure. Participants then completed 10 practice trials with fully revealed reward structures in which they could learn the statistics of the environment from experience. Next, participants answered a quiz about the range of rewards at the first step and the third step, the cost of clicking, and their bonus.

We recruited 69 participants on Amazon Mechanical Turk; 16 of them (23%) were excluded for either never clicking during the training block or incorrectly answering more than one of the four quiz questions. Each participant received a base pay of $0.50 and a performance dependent bonus that was proportional to their final score in the game (avg. bonus: $1.84 ± 0.81) for about 16.5 minutes of work on average.

Results

Participants achieved average scores of 31.3 ± 0.7 and 25.1 ± 1.1 in the increasing and decreasing conditions respectively, about 80.2% of the average score achieved by the optimal planning strategy in both cases. As in Experiment 1, participants' moves were optimal relative to the information they had uncovered on more than 95% of trials in both conditions.

Model comparisons We found that our optimal planning model (BIC = 21203) explained our participants' click sequences in the two conditions substantially better than the directed cognition model (BIC = 25354), the best-first search model (BIC = 29122), and the random model (BIC = 29313). While the optimal model and the directed cognition model correctly predicted forward versus backward planning, none of the classic models of planning can capture people's backward planning in the increasing variance condition. This highlights the advantage of having a general theory of how people's cognitive strategies are shaped by the structure of the environment and cognitive constraints.

Qualitative predictions As predicted by our resource-rational analysis, participants engaged in forward planning when the variance of the reward distribution was decreasing and backward planning when it was increasing. Concretely, in the condition with outwardly increasing variance the first click inspected a potential end state 95.0% of the time compared to 0.3% in the condition with decreasing variance ($\chi^2(1) = 520.5, p < .0001$). Conversely, in the decreasing variance condition 99.7% of the first clicks inspected one of the immediate rewards compared to only 4.1% in the increasing variance condition ($\chi^2(1) = 478.3, p < .0001$). Furthermore, when the variance increased outwardly, then only 13.9% of participants inspected any of the rewards at steps 1 or 2 before they had inspected all potential end states. Like-

wise, when the variance decreased outwardly, 86.1% of participants' second clicks also inspected an immediate reward unless the first click observed the largest possible reward in which case 69.3% of them stopped planning as predicted. Like the optimal strategy, participants in the increasing variance condition stopped 81.6% of the time they discovered a terminal state with the highest possible reward.

Conclusion

The assumption that humans are well-adapted to their environment (Anderson, 1990) greatly constrains the space of behavioral models a scientist might consider, and has thus facilitated rapid progress in many psychological domains. However, many of the problems people face are as much a product of their own computational constraints as they are a product of the external environment (Simon, 1956). By modeling these constraints as part of the problem humans must solve, we can apply the rationality assumption to a wider set of psychological phenomena.

Here, we have shown that a resource-rational analysis predicts people's planning strategies more accurately than previous models of planning, and captures how people's planning strategies depend on the structure of the external environment. This finding is congruent with recent evidence for a metacognitive reinforcement learning mechanism that makes people's cognitive strategies increasingly more resource-rational (Krueger, Lieder, & Griffiths, 2017; Lieder & Griffiths, 2017). Follow-up experiments will investigate whether the resource-rational planning strategies discovered in this work can also predict human behavior in more naturalistic sequential decision problems without the artificial constraints of our process-tracing paradigm.

One limitation of the resource-rational analysis presented here is that we have approximated bounded-optimality by rational metareasoning which assumes that the agent can determine the optimal computations at no cost (Russell, 1997). Future work replacing the optimal strategy computed by rational metareasoning with the bounded-optimal strategy computed by optimizing over implementable production systems might thus be able to explain some of the apparent sub-optimalities of human planning identified in this work.

Our study illustrates the potential of resource-rational analysis for elucidating people's cognitive strategies and understanding why they are used. Our findings suggest that this approach can make valuable contributions to the debate about human rationality by enabling a quantitative assessment of people's cognitive strategies against realistic normative standards and a fine-grained characterization of when and how they deviate from bounded-optimal information processing. Extending this approach to increasingly more realistic problems, including planning tasks where cognitive operations reveal the causal structure of the environment, is an important direction for future work.

Acknowledgement This work was supported by grant number ONR MURI N00014-13-1-0341 and a grant from the Templeton World Charity Foundation.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Psychology Press.

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.*

De Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Grouton.

Gabaix, X., Laibson, D., Moloche, G., & Weinberg, S. (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *The American Economic Review*, *96*(4), 1043–1068.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, *7*(2), 217–229.

Hay, N., Russell, S., Tolpin, D., & Shimony, S. (2012). Selecting Computations: Theory and Applications. In N. de Freitas & K. Murphy (Eds.), *Proceedings of the 28th conference on uncertainty in artificial intelligence.* Corvallis: AUAI Press.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, *8*(3), e1002410.

Huys, Q. J., 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.

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

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

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1–32.

Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2017). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*(1), 322–349.

Newell, A., & Simon, H. (1956). The logic theory machine–a complex information processing system. *IRE Transactions on information theory*, *2*(3), 61–79.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, UK: Cambridge University Press.

Russell, S. J. (1997). Rationality and intelligence. *Artificial intelligence*, *94*(1-2), 57–77.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, *63*(2), 129.

Simon, H. A. (1982). *Models of bounded rationality: Empirically grounded economic reason* (Vol. 3). Cambridge, MA: MIT press.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of experimental psychology: general*, *127*(2), 161.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1) (No. 1). Cambridge, MA: MIT press.