Exploration With a Finite Brain

Marcel Binz [1] **Eric Schulz** [1]

Abstract

Equipping artificial agents with useful exploration mechanisms remains a challenge to this day. Humans, on the other hand, seem to manage the trade-off between exploration and exploitation effortlessly. In the present article, we put forward the hypothesis that they accomplish this by making optimal use of limited computational resources. We study this hypothesis by meta-learning reinforcement learning algorithms that sacrifice performance for a shorter description length. The emerging class of models captures human exploration behavior better than previously considered approaches, such as Boltzmann exploration, upper confidence bound algorithms, and Thompson sampling. We additionally demonstrate that changes in description length produce the intended effects: reducing description length captures the behavior of brain-lesioned patients while increasing it echoes cognitive development during adolescence.

1. Introduction

Knowing how to efficiently balance between exploring unfamiliar parts of an environment and exploiting currently available knowledge is an essential ingredient of any intelligent agent. In theory, it is possible to obtain a Bayes-optimal solution to this exploration-exploitation dilemma by solving an augmented problem referred to as Bayes-adaptive Markov decision process (BAMDP, Duff, 2003). However, BAMDPs are in general intractable to solve as analytical solutions are only available for a few special cases (Gittins, 1979). The intractability of this problem prompted researchers to develop a body of heuristic solutions (Auer et al., 2002; Kaufmann et al., 2012; Russo et al., 2017; Russo & Van Roy, 2014). These heuristics can be divided into two broad categories: directed and random exploration strategies (Wilson et al., 2014; Schulz & Gershman, 2019). Directed exploration strategies provide bonus rewards that encourage the agent to visit parts of the environment that ought to be explored, whereas random exploration strategies inject some form of stochasticity into the policy.

Having a vast amount of existing exploration strategies leads to the question: which of them should we use when building artificial agents? To answer this question, we may take inspiration from how people approach the exploration-exploitation dilemma. Human exploration has been studied extensively in multi-armed bandit problems (Mehlhorn et al., 2015; Wilson et al., 2021; Brändle et al., 2021). Prior work indicates that people substantially deviate from the Bayes-optimal strategy even for the simplest bandit problems (Steyvers et al., 2009; Zhang & Angela, 2013; Binz & Endres, 2019). However, they use uncertainty estimates to intelligently guide their choices through a combination of both directed and random exploration (Wilson et al., 2014; Gershman, 2018). The question of when and why individuals rely on a particular exploration strategy has been an under-explored avenue so far.

We take the first steps towards answering these questions in the present article by looking at human exploration from a resource-rational perspective (Gershman et al., 2015; Lieder & Griffiths, 2020; Binz et al., 2022). More specifically, we investigate the hypothesis that people solve the exploration-exploitation dilemma by making optimal use of limited computational resources. To test this hypothesis, we devise a family of resource-rational reinforcement learning algorithms by combining ideas from meta-learning (Bengio et al., 1991; Schmidhuber et al., 1996) and information theory (Hinton & Van Camp, 1993). The resulting model – which we call RL^3 – implements a free-standing reinforcement learning algorithm that achieves optimal behavior subject to the constraint that it can be implemented with a given number of bits.

We demonstrate that RL^3 captures many aspects of human exploration by reanalyzing data from three previously conducted psychological studies. First, we show that it explains human choices in a two-armed bandit task on both a qualitative and quantitative level. In this setting, RL^3 discovers a set of diverse exploration strategies, allowing us to account for individual differences in human decision-making. It furthermore provides a better fit to human data than traditional

[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany. Correspondence to: Marcel Binz <marcel.binz@tuebingen.mpg.de>.

Preprint. Under Review.

approaches, such as Thompson sampling (Thompson, 1933), upper confidence bound (UCB) algorithms (Kaufmann et al., 2012), and mixtures thereof (Gershman, 2018). We then verify that the manipulation of computational resources in this class of models matches the manipulation of resources in human subjects in two different contexts: (1) Lowering the description length of RL^3 leads to decision-making defects that have been reported in patients with brain lesions. (2) Increasing the description length of RL^3 aligns with the trajectory of exploration strategies that children go through during their cognitive development. Taken together, these results enrich our understanding of human exploration and provide insights into how to improve the exploratory capabilities of artificial agents.

2. Methods

In this section, we demonstrate how to devise a resource-rational reinforcement learning algorithm via meta-learning. We first describe the general problem setting and its optimal solution, followed by a brief summary of the meta-reinforcement learning framework. We then show how to augment the standard meta-reinforcement learning objective with an information-theoretic constraint, allowing us to construct reinforcement learning algorithms that trade-off performance against the number of bits required to implement them.

2.1. Notation and Preliminaries

Each task considered in this article can be interpreted as a multi-armed bandit problem. In a k-armed bandit problem, an agent repeatedly interacts with k slot machines. These slot machines, also known as arms, are associated with a reward distribution $p(r_t|a_t, \omega)$ with unknown parameters ω. In each time-step, the agent selects an arm a_t and is provided a reward r_t based on the associated reward distribution.

The goal of an agent is to select arms such that the sum of rewards over a finite horizon H is maximized. We focus on the Bayesian setting, in which the agent additionally has access to a prior distribution $p(\omega)$ over bandit problems that it may encounter. This setting enables the agent to update its prior beliefs over reward functions after it has observed a history of observations $h_t := (a_1, r_1, \ldots, a_{t-1}, r_{t-1})$ through Bayes' rule:

$$p(\omega|h_t) \propto p(\omega) \prod_{m=1}^{t-1} p(r_m|a_m, \omega) \tag{1}$$

The policy that optimally trades-off exploration and exploitation can be obtained by reasoning how the agent's beliefs about reward functions evolve with future observations (Martin, 1967). Formally, this is accomplished by transforming the original bandit problem into a corresponding BAMDP defined by the tuple $(\mathcal{H}, \mathcal{A}, H, T, R)$. In this augmented problem, $\mathcal{H}$ represents the set of all possible histories, while $\mathcal{A}$ and H correspond to the action space and the horizon of the original bandit problem. The transition probabilities T and reward function R are given by:

$$T(h_{t+1}|a_t, h_t) = p(r_t|a_t, h_t)\delta\left[h_{t+1} = (h_t, a_t, r_t)\right] \tag{2}$$

$$R(a_t, h_t) = \mathbb{E}_{p(r_t|a_t,h_t)}\left[r_t\right] \tag{3}$$

with the marginal reward probabilities:

$$p(r_t|a_t, h_t) = \int p(r_t|a_t, \omega)p(\omega|h_t)d\omega \tag{4}$$

The policy that maximizes the sum of rewards in the BAMDP corresponds to the Bayes-optimal policy for the original bandit problem.

2.2. Meta-Reinforcement Learning

While the BAMDP formalism provides a precise recipe to derive the Bayes-optimal policy, finding an analytical expression of this policy is typically not possible. Recent work on meta-reinforcement learning, however, has shown that it is possible to learn an approximation to the Bayes-optimal policy (Wang et al., 2016; Ortega et al., 2019; Zintgraf et al., 2019). Duan et al. (2016) refer to this approach as RL^2 because it uses a traditional reinforcement learning algorithm to meta-learn another free-standing reinforcement learning algorithm.

RL^2 parametrizes the to-be-learned reinforcement learning algorithm with a general-purpose function approximator. Typically, this function approximator takes the form of a recurrent neural network that receives the last action and reward as inputs, uses them to update its hidden state, and produces a policy that is conditioned on the new hidden state. Let $\mathbf{W}$ denote the parameters of this recurrent neural network. In an outer-loop meta-learning process, the network is then trained on the prior distribution over bandit problems $p(\omega)$ to find the history-dependent policy $\pi(a_t|h_t, \mathbf{W})$ that maximizes the sum of obtained rewards. If the meta-learning procedure has successfully converged to its optimum, RL^2 implements a free-standing reinforcement learning algorithm that approximates the Bayes-optimal policy. Importantly, learning at this stage is fully implemented through the forward dynamics of the recurrent neural network and no further updating of its parameters is required.

2.3. RL^3

We transform RL^2 into a resource-rational algorithm by augmenting its meta-learning objective with an information-theoretic constraint and refer to this resource-limited variant as RL^3. More precisely, RL^3 is obtained by solving the

following optimization problem:

$$\max_{\mathbf{\Lambda}} \mathbb{E}_{q(\mathbf{W}|\mathbf{\Lambda})p(\omega)\prod p(r_t|a_t,\omega)\pi(a_t|h_t,\mathbf{W})}\left[\sum_{t=1}^{H} r_t\right]$$
$$\text{s.t. KL}\left[q(\mathbf{W}|\mathbf{\Lambda})||p(\mathbf{W})\right] \leq C \tag{5}$$

The first component of Equation 5 corresponds to the standard meta-reinforcement learning objective, while the second component ensures that the Kullback–Leibler (KL) divergence between a stochastic parameter encoding $q(\mathbf{W}|\mathbf{\Lambda})$ and a prior $p(\mathbf{W})$ remains smaller than some constant C. The KL term can be interpreted as the description length of neural network parameters, i.e., the number of bits required to store them.[1] RL^3 therefore optimally trades-off performance against the number of bits required to implement the emerging reinforcement learning algorithm. It furthermore approximates the Bayes-optimal policy as C goes to infinity under the same conditions as RL^2 (namely a sufficiently expressive function approximator and successful optimization to the global optimum). Note that the objective from Equation 5 can also be motivated by a PAC-Bayes bound on generalization performance to unseen tasks (Yin et al., 2019; Rothfuss et al., 2020; Jose & Simeone, 2020). While we focus on the resource-rational interpretation in the present article, we believe that both of these perspectives are complementary.

In practice, we solve a sample-based approximation of the optimization problem in Equation 5 using a standard on-policy actor-critic procedure (Mnih et al., 2016; Wu et al., 2017). We rely on a dual gradient ascent procedure (Haarnoja et al., 2018) to ensure that the constraint is satisfied. Appendix A contains a complete description of the network architecture, choices of prior and encoding distribution, and the meta-learning procedure. For the purpose of this article, we are only interested in the properties of the fully-converged model and not in what happens during meta-learning.

3. Individual Differences in Exploration

In an initial step, we are going to investigate how the exploration strategies implemented in RL^3 change when manipulating its description length. We are then going to test whether the set of emerging strategies describes human behavior well on a quantitative level. When reanalyzing data from a two-armed bandit benchmark (Gershman, 2018), we find that RL^3 beats previously considered algorithms in terms of fitting human behavior by a large margin.

Experimental Design: The behavioral data-set of Gershman (2018) contains records of 44 participants who each played 20 two-armed bandit problems with an episode length of $H = 10$. The mean reward for each arm a was drawn from $p(\omega_a) = \mathcal{N}(0, 10)$ at the beginning of the task and the reward in each time-step from $p(r_t|a_t, \omega) = \mathcal{N}(\omega_{a_t}, 1)$.

Analysis: To analyze the set of emerging exploration strategies, we adopted a method proposed by Gershman (2018). He assumed that an agent (either human or machine) uses Bayes' rule as described in Equation 1 to update its beliefs over unobserved parameters. If prior and reward are both normally distributed, the posterior will also be normally distributed and the corresponding updating rule is given by the Kalman filtering equations. Let $p(\omega_a|h_t) = \mathcal{N}(\mu_{a,t}, \sigma_{a,t})$ be the posterior distribution at time-step t. Based on the parameters of this posterior distribution, he then defined the following probit regression model:

$$p(A_t = 0|h_t, \mathbf{w}) = \mathbf{\Phi}\left(\mathbf{w}_1 \mathrm{V}_t + \mathbf{w}_2 \mathrm{RU}_t + \mathbf{w}_3 \frac{\mathrm{V}_t}{\mathrm{TU}_t}\right)$$
$$\mathrm{V}_t = \mu_{0,t} - \mu_{1,t}$$
$$\mathrm{RU}_t = \sigma_{0,t} - \sigma_{1,t} \tag{6}$$
$$\mathrm{TU}_t = \sqrt{\sigma_{0,t}^2 + \sigma_{1,t}^2}$$

with $\mathbf{\Phi}$ denoting the cumulative distribution function of a standard normal distribution. Equation 6 is also referred to as the hybrid model as it contains several known exploration strategies as special cases. We can recover a Boltzmann-like exploration strategy for $\mathbf{w} = [\mathbf{w}_1, 0, 0]$, a variant of the UCB algorithm for $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, 0]$, and Thompson sampling for $\mathbf{w} = [0, 0, 1]$.

Fitting the coefficients of the hybrid model to behavioral data allows us to inspect how much a given agent relied on a particular exploration strategy. Previously, Gershman (2018) has applied this form of analysis to human data, which revealed that people rely on a mixture of directed and random exploration. In this article, we extend this approach to artificial data generated by RL^3. In addition, we use the hybrid model, and the previously discussed sub-models, as baselines against which we compare our proposed model.

Results: We trained RL^3 with a targeted description length of $\{1, 2, \ldots, 10000\}$ nats on the same distribution used in the original experimental study. Figure C1 (a) confirms that performance improves as description length is increased. Figure C1 (b) verifies that our models achieved their targeted description length.

We first examined how the description length of RL^3 influences its exploration behavior using the previously described probit regression analysis. Figure 1 (a) illustrates the results of this analysis. We find that RL^3 implements a Boltzmann-like exploration strategy for description lengths between

[1] The desired coding length can, for example, be achieved using bits-back coding (Hinton & Van Camp, 1993) or minimal random coding (Havasi et al., 2018).

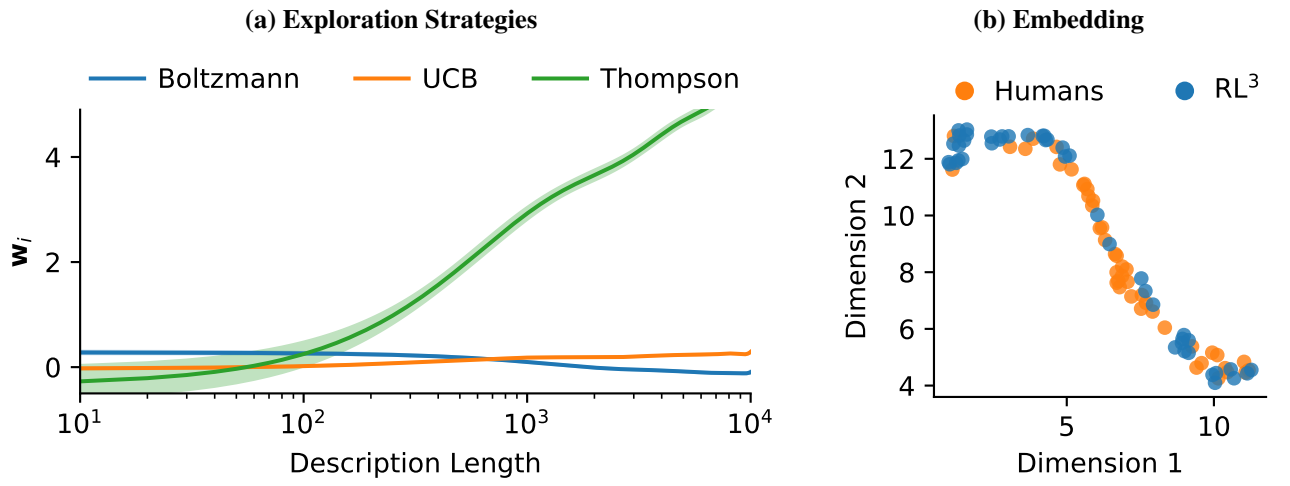

Figure 1. Illustration of exploration strategies realized by RL^3. (a) Probit regression coefficients obtained by fitting the hybrid model to data simulated by RL^3 with varying description lengths. Depicted on a logarithmic scale. (b) UMAP embeddings of probit regression coefficients for RL^3 and human participants.

1 and 100 nats. Note that behavior at this stage is quite noisy as indicated by the small probit regression weights. Beginning from 100 nats, we observe a rise of the factor corresponding to Thompson sampling. This factor continues to rise until the limit of 10000 nats. Between 100 and 1000 nats, we additionally find minor influences of a Boltzmann-like exploration strategy. For a description length of 1000 nats and larger, Boltzmann-like exploration diminishes and is replaced with minor influences of a UCB-based strategy.

Having established that different exploration styles emerge in RL^3 depending on its description length, we next set out to test how well it explains human choices. In order to do so, we conducted a Bayesian model comparison (Bishop, 2006). A detailed summary of our comparison procedure is provided in Appendix B. We used the Bayesian information criterion (BIC, Schwarz, 1978) as an approximation to the model evidence. The resulting BIC values are illustrated in Figure 2 (a). We find that the BIC value for RL^3 is substantially lower compared to that of the hybrid model (5562.63 against 6158.91) when aggregated across all participants; all other models provide a less adequate fit to human choices. The majority of participants ($n = 32$) was best described by RL^3 and the protected exceedance probability (PXP), which measures the probability that a particular model is the most frequent within a set of alternatives (Rigoux et al., 2014), also favored RL^3 decisively (PXP ≈ 1). We provide a detailed illustration of the posterior probabilities for each model and participant in Figure 2 (b).

Finally, we compared the probit regression coefficients of human participants to the ones of RL^3. Figure 1 (b) shows a two-dimensional UMAP embedding (McInnes et al., 2018) of these coefficients. The figure reveals a set of diverse exploration strategies within the human population and confirms that RL^3 captures the overall variability of human strategies appropriately.

4. Manipulating Computational Resources

RL^3 also makes precise predictions about what should happen if computational resources are actively manipulated. Do these predictions align with the actual behavior of people? Providing answers to this question is non-trivial because we cannot simply ask a person to use an algorithm with a shorter or longer description length. There are, however, at least two types of studies that can provide insights. The first type includes lesion studies that compare the behavior of healthy participants to that of participants suffering from brain damage, while the second includes developmental studies that investigate how behavior evolves during cognitive development. We next take a look at an example of each of them and demonstrate that RL^3 reproduces their key findings.

4.1. Damage to Ventromedial Prefrontal Cortex

There has been a long history of analyzing people with brain lesions in cognitive neuroscience (Damasio, 1989). We focused on a particular study conducted by Bechara et al. (1994) for the purpose of this article and predicted that reducing the description length of RL^3 should correspond to the behavior of brain-lesioned patients.

Experimental Design: To probe decision-making in clinical populations, Bechara et al. (1994) introduced a novel experimental paradigm called the Iowa Gambling Task (IGT). The IGT involves 100 choices in a single four-armed ban-

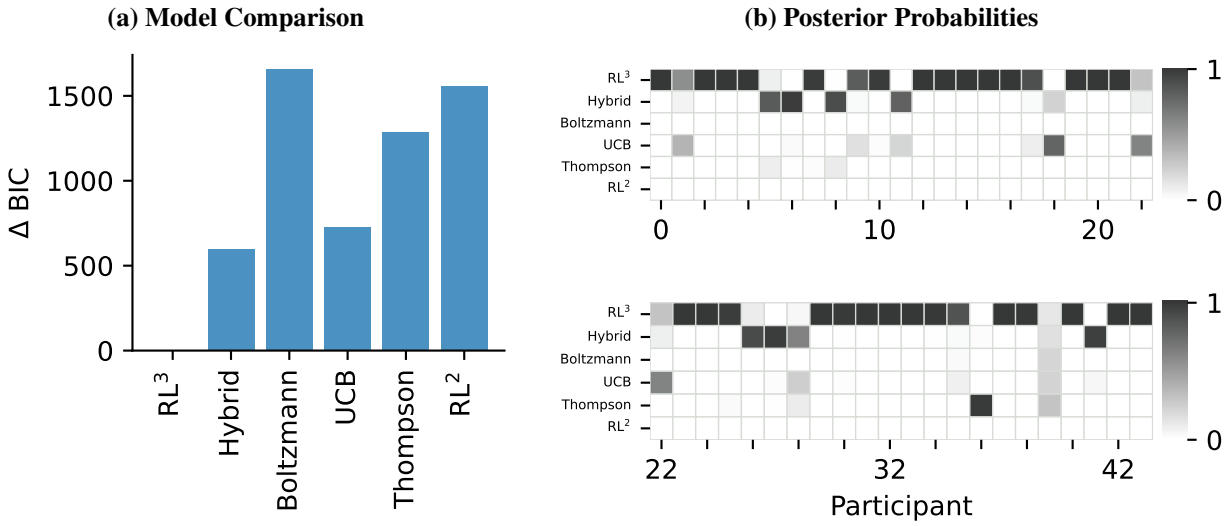

Figure 2. Model comparison results on the two-armed bandit data from Gershman (2018). (a) Bayesian information criterion (BIC) values for the aggregated data of all participants. Lower values correspond to a better fit to human behavior. (b) Posterior probabilities for each model and participant. Higher values correspond to a better fit to human behavior.

dit problem.[2] Two of the arms are high-risk arms, while the other two are low-risk arms. High-risk arms result in a constant positive reward of 100 but have a chance to yield a noisy penalty with an expected value of 125. Low-risk arms result in a constant positive reward of 50 but have a chance to yield a noisy penalty with an expected value of 25. A complete list of trials is printed in Table D1. Bechara et al. (1994) used the IGT to compare the decision-making of healthy participants to that of participants with ventromedial prefrontal cortex (vmPFC) damage.

Analysis: We analyzed the proportion of selected low- and high-risk arms across the entire experiment. High-risk arms cause an average loss of 25 points per trial, while low-risk arms provide an average payoff of 25 points per trial. We should therefore expect an agent to select the superior low-risk arms with higher frequency. Healthy participants are indeed able to learn about the structure of the task and will after a while start to sample the superior low-risk arms. Participants with vmPFC damage, however, continue to sample to the inferior high-risk as illustrated in Figure 3 (a). This pattern is striking because performance in these subjects remains worse than chance regardless of how often they interact with the task.

Results: RL^3 requires a distribution over bandit problems for meta-learning, but participants in the IGT only encountered a single bandit task. Therefore, we cannot directly use the task of the original study for meta-learning as we have done in the previous example. We instead decided to construct a distribution over bandit problems that maintains the characteristics of the IGT (a constant positive reward component plus a sparse and noisy negative reward component). While there are many distributions that fulfill these criteria, we tried to implement them in a minimalist fashion:

- The positive component was independently sampled for each arm from a uniform distribution with a minimum value of 0 and a maximum value of 150.
- The mean across all trials of the negative component was also sampled from a uniform distribution with a minimum value of 0 and a maximum value of 150.
- The negative component had an occurrence probability sampled randomly from a uniform distribution with a minimum value of 0.05 and a maximum value of 0.95.
- We furthermore added additive noise sampled from a zero-mean normal distribution with a standard deviation of 10 to the negative component in each time-step.

We trained RL^3 with a targeted description length of $\{100, 200, \ldots, 10000\}$ nats on the previously described distribution. Figure D1 (a) confirms that performance improves as description length is increased. Figure D1 (b) verifies that our models achieved their targeted description length.

When tested on the IGT, we find that RL^3 replicates the pattern reported by Bechara et al. (1994). Models with a high description length successfully solve the task by selecting low-risk arms in the majority of time-steps. If description

[2]The task is typically phrased in form of a repeated selection of cards from one out four decks. This formulation is mathematically equivalent to the four-armed bandit description used in our article.

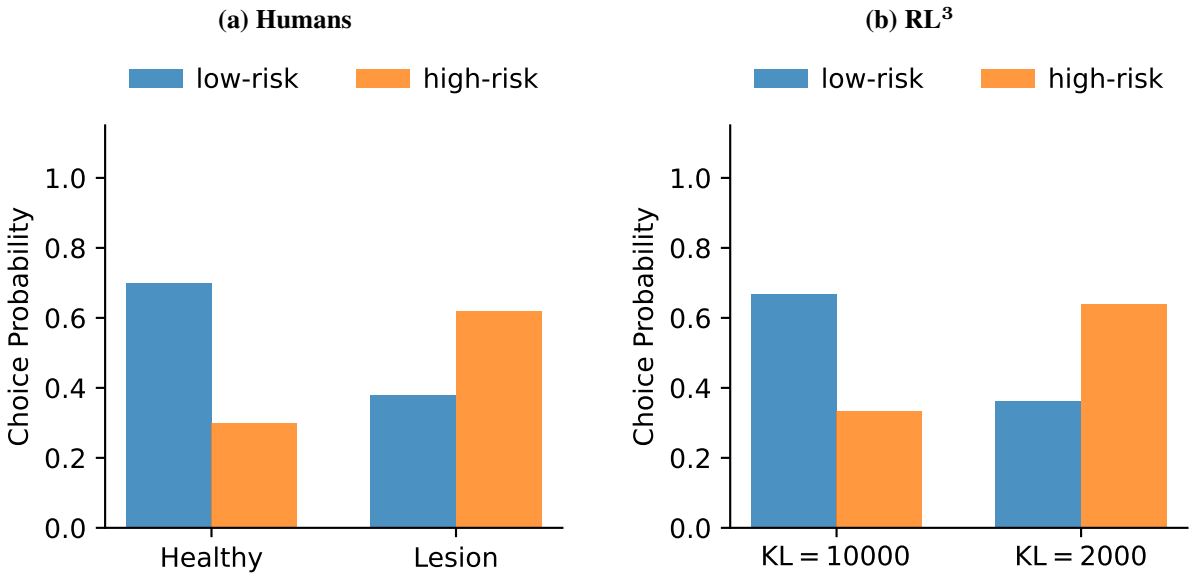

Figure 3. Probability of selecting low- and high-risks arms in the Iowa Gambling Task. (a) Human data taken from Bechara et al. (1994). The probability of selecting an inferior high-risk arm is increased in participants with vmPFC damage. (b) Data simulated from RL^3 with large and small description length. The probability of selecting an inferior high-risk arm is increased for models with fewer bits, mirroring the results of the original study.

length is however sufficiently reduced, RL^3 predominately samples high-risk arms. We illustrate this behavior for two example models in Figure 3 (b). Figure D2 provides a more detailed picture of how description length mediates choice behavior.

In summary, our analysis sheds light on why brain-lesioned patients display below-chance performance in the IGT. Intuitively, any resource-limited agent must primarily devote its computational resources to things that are easy to estimate. In the IGT, the deterministic positive component is easier to estimate than the noisy negative component. An agent with significantly restricted resources will thus focus on the positive component while ignoring the negative. In turn, the agent will assign higher estimated payoffs to the inferior high-risk arms and therefore select them more frequently. We found that RL^3 implements this behavior and that reducing its description length captured participants with lesioned vmPFC.

4.2. Developmental Trajectories

People are not born with fully-developed cognitive abilities but instead develop them during their lifetime. In this section, we tested whether increasing the description length of RL^3 matches the behavioral trajectories of people as they grow up. To test this hypothesis, we reanalyzed data collected by Somerville et al. (2017), who studied changes in exploration behavior between early adolescence and adulthood.

Experimental Design: In their study, Somerville et al. (2017) made use of an experimental paradigm known as the horizon task (Wilson et al., 2014). Each task was based on a two-armed bandit problem and involves four forced-choice trials, followed by either one or six free-choice trials. Participants were aware of the number of remaining choices and could use this information to guide their behavior. The mean reward of one of the arms was drawn randomly from $[40, 60]$, while the mean reward for the other was determined by sampling the difference to the first arm from $[4, 8, 12, 20, 30]$. The arrangement of arms and the sign of the difference was randomized. In each time-step, the observed reward was sampled from a normal distribution with the corresponding mean value and a standard deviation of 8. The addition of forced-choice trials allowed to control the amount of information that was available to participants. They either provided an equal amount of information for both arms (i.e., two observations each) or an unequal amount of information (i.e., a single observation from one arm, three from the other). In total, Somerville et al. (2017) collected data for 147 participants between the ages 12.08 and 28, completing 160 bandit tasks each.

Analysis: Following Somerville et al. (2017), we used the decision in the first free-choice trial to distinguish between different types of exploration. In the unequal information condition, a choice was classified as directed exploration if it corresponded to the option that was observed fewer times during the forced-choice trials. In the equal information condition, a choice was classified as random exploration

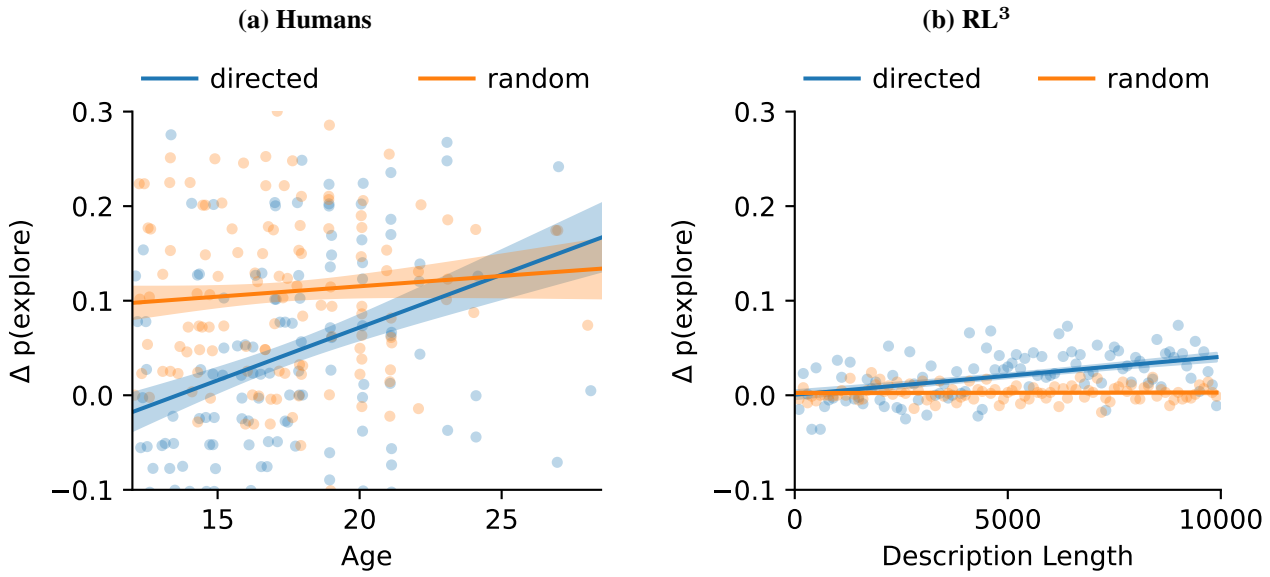

Figure 4. Illustration of strategic directed and random exploration in the horizon task. (a) Human data from Somerville et al. (2017). During adolescence, people start to engage more in strategic directed exploration, whereas strategic random exploration remains constant over time. (b) Data simulated from RL^3 with varying description lengths. Like in the human data, we observe an increase in strategic directed exploration, but no change in strategic random exploration.

if it corresponded to the option with the lower estimated mean. We refer to an exploration behavior as strategic if it occurs more frequently in the long horizon tasks compared to the short horizon tasks. Somerville et al. (2017) found – as shown in Figure 4 (a) – that strategic directed exploration emerges during adolescence, whereas strategic random exploration remains constant over time.

To quantify these effects, they fitted two independent linear regression models, using the probability of engaging in directed and random exploration as dependent variables. Both models used age, the corresponding horizon, and the interaction between the two as regressors. They found a significant effect of horizon in both conditions, indicating that participants engaged more in both directed and random exploration in tasks with a longer horizon. Furthermore, they found a significant interaction effect between horizon and age for directed exploration but not for random exploration. This confirmed that strategic directed exploration increases during cognitive development, while strategic random exploration is age-invariant.

Results: We trained RL^3 with a targeted description length of $\{100, 200, \ldots, 10000\}$ nats on the same distribution used in the original experimental study. Figure E1 (a) confirms that performance improves as description length is increased. Figure E1 (b) verifies that our models achieved their targeted description length.

Figure 4 (b) visualizes how strategic directed and random exploration change as the description length of RL^3 increases. Matching the main result of the experimental study, we find that strategic directed exploration increases with description length, while strategic random exploration remains unaffected. We repeated the previously described regression analysis on data simulated by RL^3 to quantify this conclusion (replacing age as a regressor with description length). The outcome of this analysis mirrored the results of the original study. We found a significant effect of horizon on both directed ($F_{1,194} = 56.50, p < 0.001, \eta^2 = 0.20$) and random exploration ($F_{1,194} = 6.80, p = 0.01, \eta^2 = 0.03$). This means that RL^3 made more exploratory decisions of both types if it was beneficial to do so. We also found a significant interaction effect between horizon and description length on directed exploration ($F_{1,194} = 17.48, p < 0.001, \eta^2 = 0.06$) but not on random exploration ($F_{1,194} = 0.02, p = 0.89$). These results confirm that description length and age have comparable qualitative effects on the development of strategic exploration.

However, when comparing the effect sizes of our analysis to those from the experimental study, we find that the interaction effect between description length and horizon on directed exploration only amounts to around half of the interaction effect between age and horizon ($\eta^2 = 0.06$ in RL^3 versus $\eta^2 = 0.115$ in the human population). We speculated that part of this difference comes from a mismatch between the distribution used to train our models and what kind of tasks people expect in the experiment. People, for instance,

might assume that task rewards are noisier than they are, which would require more exploratory choices, and, in turn, lead to stronger effects. We tested this hypothesis by retraining RL^3 on the same distribution but with the standard deviation of the reward noise increased by 50%. This modification increased the effect size of the interaction effect on directed exploration to $\eta^2 = 0.08$ while keeping all other effects intact. Even though this is an improvement, it does not close the gap entirely.[3] It, therefore, might be plausible that additional factors contribute to the development of strategic directed exploration during adolescence.

5. General Discussion

The exploration-exploitation dilemma is one of the core challenges in reinforcement learning. How do humans arbitrate between exploration and exploitation, and which kind of exploration strategies do they engage in? We have put forward the hypothesis that people tackle this problem in a resource-rational manner. To test this hypothesis, we proposed a method for meta-learning reinforcement learning algorithms with limited description length. The resulting class of models – which we refer to as RL^3 – makes precise predictions about how people make decisions. We have put these predictions to a rigorous test by comparing our model to data from three psychological studies. RL^3 displayed key elements of human decision-making in all three of them:

1. It captured individual differences of human exploration in a two-armed bandit task on both a qualitative and quantitative level.
2. Reducing its description length aligned with decision-making in brain-lesioned patients.
3. Increasing its description length reflected changes in exploration behavior observed during cognitive development.

In summary, our results demonstrate that it is possible to meta-learn resource-rational reinforcement learning algorithms and indicate that human exploration is well-characterized by these very algorithms.

5.1. Limitations and Future Work

In our article, we have focused on comparing RL^3 to human exploration in the simple multi-armed bandit setting. In the real world, however, people face much more sophisticated challenges that call for a richer repertoire of exploration strategies (Schulz et al., 2019; Brändle et al., 2021). This criticism is not necessarily a shortcoming of the proposed model, which could in principle be applied to more complex tasks, but rather one regarding the experimental research, which has predominately focused on multi-armed bandit problems. In future work, we intend to develop new experimental paradigms that allow us to compare RL^3 against human behavior in more complex settings.

We have also usually assumed that the experimental distribution matches the prior distribution over tasks expected by our agents. In reality, however, this assumption may be violated as descriptions of experiments typically only provide partial information about which kind of tasks to expect. We believe that this mismatch might explain the partial discrepancies in effect sizes we observed between the developmental trajectories of humans and RL^3. It is, for example, conceivable that elementary and high school students experience very different problems in their daily life, and consequently also expect very different problems in an experimental setting. Extracting which distribution over tasks a subject expects is a non-trivial question and an interesting avenue for future research.

RL^3 places a constraint on a particular type of computational resource: the description length of the reinforcement learning algorithm in use. People, on the other hand, are subject to a variety of additional computational constraints. They can, for instance, only run algorithms with finite computation time or only store a restricted amount of chunks in their short-term memory. Future work should aim to unify all of these constraints in a common framework.

5.2. Conclusion

Many applications could benefit from the availability of human-like agents. Having access to such agents may be especially valuable in cooperative self-play scenarios, where training with them is crucial for successful coordination with people (Carroll et al., 2019; Strouse et al., 2021). The traditional path towards constructing agents that learn and think like people is to take inspiration from the cognitive processes of the human mind and incorporate them into existing systems (Lake et al., 2017). In this article, we have pursued a different approach. Instead of hard-coding cognitive processes directly into our agents, we have identified two computational principles – meta-learning and resource rationality – that give rise to many aspects of human behavior. The presented approach is very general, easy to adapt to new domains, and can be scaled to more complex problem settings without major modifications. Finally, we want to emphasize that low description lengths might not only be a biological necessity, but also a feature. Implementing an algorithm in just a few bits acts as a strong form of regularization and could, in turn, produce exploration strategies that are applicable across domains. Hence, we believe that constructing artificial systems with such constraints could lead us towards more generally capable agents.

[3] Further increasing the standard deviation of the reward noise did not lead to any noticeable qualitative or quantitative changes.

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Bechara, A., Damasio, A. R., Damasio, H., and Anderson, S. W. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50(1-3): 7–15, 1994.

Bengio, Y., Bengio, S., and Cloutier, J. Learning a synaptic learning rule. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume 2, pp. 969–vol. IEEE, 1991.

Binz, M. and Endres, D. Where do heuristics come from? In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society*, pp. 1402–1408, 2019.

Binz, M., Gershman, S. J., Schulz, E., and Endres, D. Heuristics from bounded meta-learned inference. *Psychological Review*, (Advance online publication), 2022.

Bishop, C. M. Machine learning and pattern recognition. *Information science and statistics. Springer, Heidelberg*, 2006.

Brändle, F., Binz, M., and Schulz, E. Exploration beyond bandits. 2021.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. *Advances in Neural Information Processing Systems*, 32:5174–5185, 2019.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Damasio, H. Lesion analysis. *Neuropsychology.*, 1989.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., and Abbeel, P. Rl ^{2}: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Duff, M. O. Optimal learning: Computational procedures for bayes-adaptive markov decision processes. 2003.

Eysenbach, B., Salakhutdinov, R. R., and Levine, S. Robust predictable control. *Advances in Neural Information Processing Systems*, 34, 2021.

Gershman, S. J. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018.

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349 (6245):273–278, 2015.

Gittins, J. C. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177, 1979.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Havasi, M., Peharz, R., and Hernández-Lobato, J. M. Minimal random code learning: Getting bits back from compressed model parameters. In *International Conference on Learning Representations*, 2018.

Hinton, G. E. and Van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13, 1993.

Jose, S. T. and Simeone, O. Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization, 2020.

Kaufmann, E., Cappé, O., and Garivier, A. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pp. 592–600, 2012.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28:2575–2583, 2015.

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

Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Lieder, F. and Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, 2020.

Martin, J. J. *Bayesian decision problems and Markov chains*. Wiley, 1967.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., and Gonzalez, C. Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, 2(3):191, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Molchanov, D., Ashukha, A., and Vetrov, D. Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning*, pp. 2498–2507. PMLR, 2017.

Ortega, P. A., Wang, J. X., Rowland, M., Genewein, T., Kurth-Nelson, Z., Pascanu, R., Heess, N., Veness, J., Pritzel, A., Sprechmann, P., et al. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. Bayesian model selection for group studies—revisited. *Neuroimage*, 84:971–985, 2014.

Rothfuss, J., Fortuin, V., and Krause, A. Pacoh: Bayes-optimal meta-learning with pac-guarantees. *arXiv preprint arXiv:2002.05551*, 2020.

Russo, D. and Van Roy, B. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pp. 1583–1591, 2014.

Russo, D., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*, 2017.

Schmidhuber, J., Zhao, J., and Wiering, M. Simple principles of metalearning. *Technical report IDSIA*, 69:1–23, 1996.

Schulz, E. and Gershman, S. J. The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, 55:7–14, 2019.

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., and Gershman, S. J. Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, 116(28):13903–13908, 2019.

Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, pp. 461–464, 1978.

Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., and Wilson, R. C. Charting the expansion of strategic exploratory behavior during adolescence. *Journal of experimental psychology: general*, 146(2):155, 2017.

Steyvers, M., Lee, M. D., and Wagenmakers, E.-J. A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53(3): 168–179, 2009.

Strouse, D., McKee, K., Botvinick, M., Hughes, E., and Everett, R. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34, 2021.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., and Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074, 2014.

Wilson, R. C., Bonawitz, E., Costa, V. D., and Ebitz, R. B. Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, 38:49–56, 2021.

Wu, Y., Mansimov, E., Grosse, R. B., Liao, S., and Ba, J. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. *Advances in neural information processing systems*, 30: 5279–5288, 2017.

Yin, M., Tucker, G., Zhou, M., Levine, S., and Finn, C. Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*, 2019.

Zhang, S. and Angela, J. Y. Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems*, pp. 2607–2615, 2013.

Zintgraf, L., Shiarlis, K., Igl, M., Schulze, S., Gal, Y., Hofmann, K., and Whiteson, S. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.

A. Meta-Learning Details

A.1. Meta-Learning Procedure

We employed a standard on-policy actor-critic procedure (Mnih et al., 2016; Wu et al., 2017) to optimize a sample-based approximation of Equation 5 using the ADAM optimizer with a learning rate of 3×10^{-4}. We simultaneously updated a dual parameter to satisfy the constraint on description length as suggested in prior work (Haarnoja et al., 2018; Eysenbach et al., 2021). Models were trained on one million batches of size 32. By the end of meta-learning, all models have converged. We additionally scaled rewards to roughly fall within the range of $[-1, 1]$ to further stabilize training. Pseudocode for the meta-learning procedure is provided in Algorithm 1.

Algorithm 1 Meta-Learning Procedure

Initialize $\mathbf{\Lambda}$ and β
for $n = 1$ **to** N **do**
  $\omega \sim p(\omega)$
  $\mathbf{W} \sim q(\mathbf{W}|\mathbf{\Lambda})$
  a_0, r_0, h_0 = model.init()
  for $t = 1$ **to** H **do**
    $h_t, V_t, \pi(a_t|h_t, \mathbf{W})$ = model.forward$(a_{t-1}, r_{t-1}, h_{t-1})$
    $a_t \sim \pi(a_t|h_t, \mathbf{W})$
    $r_t \sim p(r_t|a_t, \omega)$
  end for
  $\mathcal{L}_{\text{dual}} \leftarrow -\beta \left(\text{KL}\left[q(\mathbf{W}|\mathbf{\Lambda})||p(\mathbf{W})\right] - C\right)$
  $\mathcal{L}_{\text{critic}} \leftarrow 0$
  $\mathcal{L}_{\text{actor}} \leftarrow 0$
  for $t = 1$ **to** H **do**
    $\mathcal{L}_{\text{critic}} \leftarrow \mathcal{L}_{\text{critic}} + \left(r_t + V_{t+1} - V_t\right)^2$
    $\mathcal{L}_{\text{actor}} \leftarrow \mathcal{L}_{\text{actor}} - \left(r_t + V_{t+1} - V_t\right) \log \pi(a_t|h_t, \mathbf{W})$
  end for
  $\mathbf{\Lambda} \leftarrow \mathbf{\Lambda} - \alpha \nabla_{\mathbf{\Lambda}} \left(H^{-1}\left(\mathcal{L}_{\text{critic}} + \mathcal{L}_{\text{actor}}\right) + \beta \text{KL}\left[q(\mathbf{W}|\mathbf{\Lambda})||p(\mathbf{W})\right]\right)$
  $\beta \leftarrow \beta - \alpha \nabla_{\beta} \mathcal{L}_{\text{dual}}$
end for

A.2. Model Architecture

The model architecture consists of a single gated recurrent unit (GRU, Cho et al., 2014) layer with a hidden size of 128. Inputs to this GRU layer correspond to the action and reward from the previous time-step. Its outputs were then transformed by two linear layers, projecting to the policy and value estimate respectively.

A.3. Prior and Encoding Distributions

The prior over network weights corresponds to a variational dropout prior (Kingma et al., 2015). The encoding distribution is parametrized by a set of independent normal distributions over network weights. We adopted the approximation suggested by Molchanov et al. (2017) to estimate the KL divergence between the encoding and prior distribution and obtained gradients with respect to $\mathbf{\Lambda}$ using the reparametrization trick (Kingma & Welling, 2013).

B. Bayesian Model Comparison

In the main text, we conducted a Bayesian model comparison to evaluate how well each model fitted the behavioral data from the two-armed bandit task. For this comparison, we computed the posterior probability that participant i with corresponding data $\mathcal{D}_i$ used model m via Bayes' theorem:

$$p(m|\mathcal{D}_i) \propto p(\mathcal{D}_i|m)p(m) \tag{7}$$

We assumed a uniform prior over models and approximated the model evidence with the Bayesian information criterion:

$$\log p(\mathcal{D}_i|m) \approx -\frac{1}{2}|\theta| \log(NH) + \max_{\theta} \sum_{n=1}^{N} \sum_{t=1}^{H} \log p(A_t^{i,n} = a_t^{i,n}|h_t^{i,n}, \theta, m) \tag{8}$$

where N is the total number of tasks, H is the number of trials within each task, and $|\theta|$ is the number of fitted parameters. We use $a_t^{i,n}$ and $h_t^{i,n}$ to denote the action chosen and the history observed by participant i in task n and trial t. The policy of our baseline models is directly given by Equation 6. For RL^2 and RL^3, we additionally assumed an ϵ-greedy error model:

$$p(A_t = 0|h_t, \theta, m) = (1 - \epsilon)\pi(A_t = 0|h_t, \theta, m) + 0.5\epsilon \tag{9}$$

Furthermore, we approximated the integral over neural network weights with a Monte Carlo approximation of $S = 10$ samples:

$$\pi(A_t = 0|h_t, \theta, m) \approx \frac{1}{S} \sum_{s=1}^{S} \pi(A_t = 0|h_t, \mathbf{W}_s, \theta) \qquad \mathbf{W}_s \sim q(\mathbf{W}|\mathbf{\Lambda}) \tag{10}$$

Table B1 specifies fitted parameters θ for each model and their corresponding search domains. We applied a simple grid search procedure over the values given in Table B1 to obtain the maximum likelihood estimate of parameters for RL^2 and RL^3. Parameters of the baseline models were optimized using a Newton-Raphson algorithm.

Table B1. Fitted parameters in each model, together with their corresponding search domains.

Model	**Parameters**	**Domain**
RL^3	ϵ	$\{0.01, 0.02, \ldots, 1\}$
	C	$\{1, 2, \ldots, 10000\}$
Hybrid	$\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$	continuous
Boltzmann	$\mathbf{w}_1$	continuous
UCB	$\mathbf{w}_1, \mathbf{w}_2$	continuous
Thompson	$\mathbf{w}_3$	continuous
RL^2	ϵ	$\{0.01, 0.02, \ldots, 1\}$

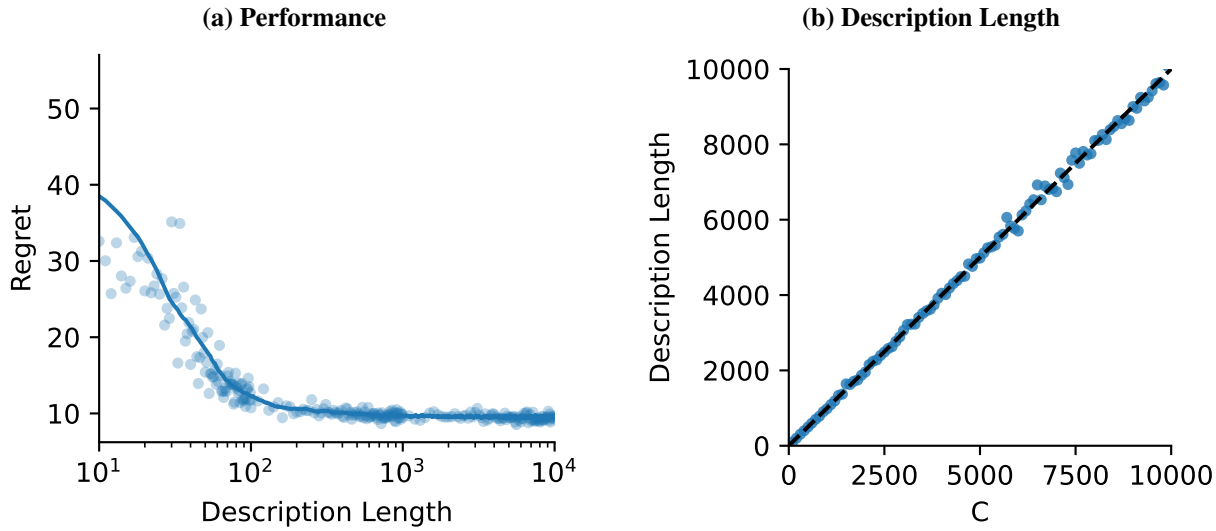

Figure C1. Performance and description lengths in the two-armed bandit task. (a) Regrets for models with varying description lengths. Depicted on a logarithmic scale to ensure comparability with Figure 1. The solid line shows a moving average over a window of fifty. (b) Targeted description lengths C plotted against description lengths of the converged models.

C. Two-Armed Bandit Task

Models were trained as described in Appendix A. Figure C1 (a) confirms that performance improves as description length is increased. Figure C1 (b) verifies that our models achieved their targeted description length.

D. Iowa Gambling Task

Models were trained as described in Appendix A. In addition, we assumed a discount factor of $\gamma = 0.9$ for this set of experiments. A geometric discount factor can be interpreted as a modification to the task dynamics such that an agent believes to reach a terminal state with probability $1 - \gamma$ (Levine, 2018). We used this interpretation of the discount factor to model that participants in the IGT are not informed about the duration of the task.

Figure D1 (a) confirms that performance improves as description length is increased. Figure D1 (b) verifies that our models achieved their targeted description length. Table D1 contains a trial-wise overview of the IGT. Figure D2 depicts the probability of selecting low- and high-risks arms for all description lengths.

Table D1. Example of ten consecutive trials in the Iowa Gambling Task. We assume ten blocks of ten trials each. The order of trials within each block is randomized. The top row for each arm shows the deterministic positive component, while the bottom row shows the noisy negative component.

	Trial									
Arm	1	2	3	4	5	6	7	8	9	10
1	+100	+100	+100	+100	+100	+100	+100	+100	+100	+100
	0	0	−150	0	−300	0	−200	0	−250	−350
2	+100	+100	+100	+100	+100	+100	+100	+100	+100	+100
	0	0	0	0	0	0	0	0	−1250	0
3	+50	+50	+50	+50	+50	+50	+50	+50	+50	+50
	0	0	−50	0	−50	0	−50	0	−50	−50
4	+50	+50	+50	+50	+50	+50	+50	+50	+50	+50
	0	0	0	0	0	0	0	0	0	−250

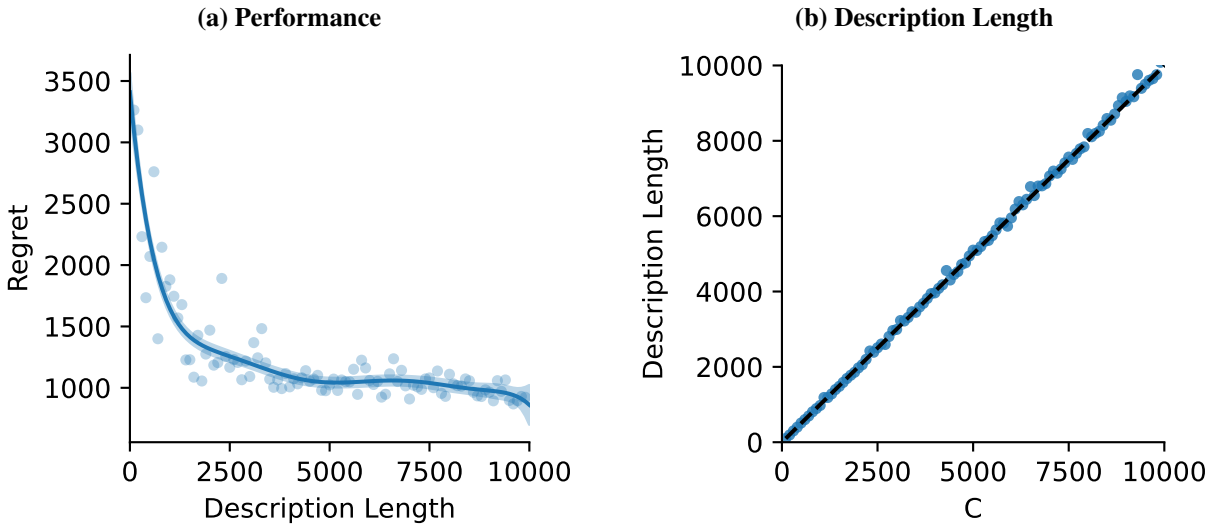

Figure D1. Performance and description lengths in the meta-learning version of the Iowa Gambling Task. (a) Regrets for models with varying description lengths. The solid line shows the mean prediction of a Bayesian polynomial regression model. Shaded contours represent the standard deviation of the mean. (b) Targeted description lengths C plotted against description lengths of the converged models.

E. Horizon Task

Models were trained as described in Appendix A. In addition, models received a binary value encoding the horizon of the current task. Like the humans in the experimental study, they could use this information to guide their exploration behavior. We unrolled the network during the forced-choice trials by providing it with the externally specified actions and rewards. We did not use the forced-choice trials to update the parameters of the network.

Figure E1 (a) confirms that performance improves as description length is increased. Figure E1 (b) verifies that our models achieved their targeted description length.

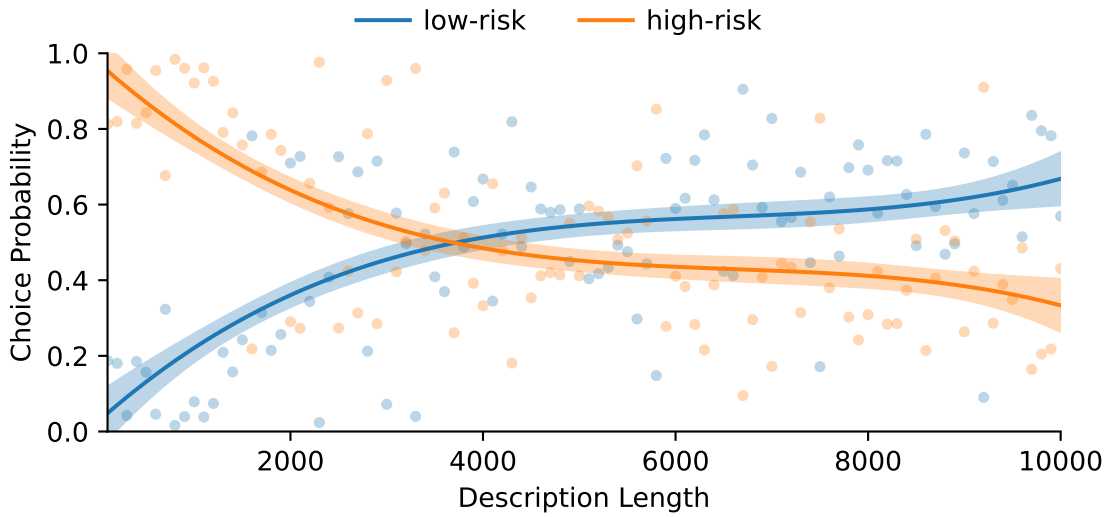

Figure D2. Probability of selecting low- and high-risks arms in the Iowa Gambling Task for all description lengths. Mean choice probabilities are illustrated as solid lines obtained by fitting a Bayesian polynomial regression model to the underlying data. Shaded contours represent the standard deviation of the mean.

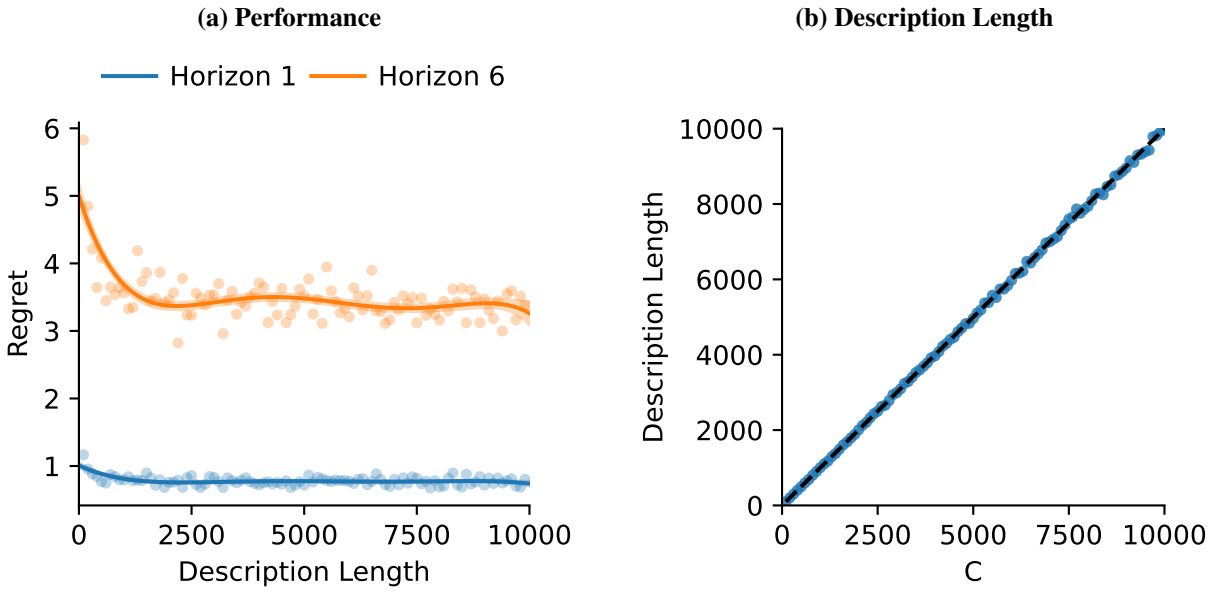

Figure E1. Performance and description lengths in the horizon task. (a) Regrets for models with varying description lengths. The solid line shows the mean prediction of a Bayesian polynomial regression model. Shaded contours represent the standard deviation of the mean. (b) Targeted description lengths C plotted against description lengths of the converged models.