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 $\beta$
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
        $\mathcal{L}_{\text{critic}} \leftarrow \mathcal{L}_{\text{critic}} + (r_t + V_{t+1} - V_t)^2$
        $\mathcal{L}_{\text{actor}} \leftarrow \mathcal{L}_{\text{actor}} - (r_t + V_{t+1} - V_t) \log \pi(a_t|h_t, \mathbf{W})$
    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 and 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 RR-RL^2, we additionally assumed an ε-greedy error model:

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

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

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

In addition to the two meta-learning models, we considered several baseline algorithms within our model comparison procedure. These include the hybrid model from Equation 6, Thompson sampling, an upper confidence bound algorithm, and a Boltzmann-like exploration strategy. Note that our baseline algorithms can be obtained by restricting a sub-set of parameters in the hybrid model to be zero. 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 RR-RL^2. 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**
RR-RL^2	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	$\{0.01, 0.02, \ldots, 1\}$

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.

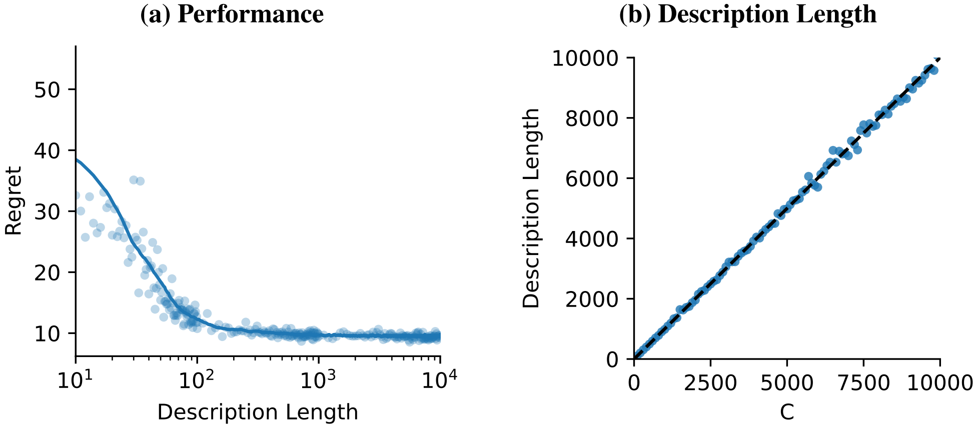


Figure C1: Performance and description lengths in the two-armed bandit task. (a) Regrets for models with varying description lengths. Depicted on a logarithmic scale to ensure comparability with Figure 1. The solid line shows a moving average over a window of fifty. (b) Targeted description lengths C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

C.1 Model Comparison Based on AIC

We complemented our Bayesian model comparison from the main text with a model comparison that relies on the Akaike information criterion (AIC, Akaike, 1974):

$$\text{AIC} = 2|\theta| - 2 \max_{\theta} \sum_{n=1}^{N} \sum_{t=1}^{H} \log p(A_t^{i,n} = a_t^{i,n} | h_t^{i,n}, \theta, m) \tag{11}$$

This measure offers a simple approximation to the out-of-sample predictive accuracy of a model [Gelman et al., 1995]. Figure C2 visualizes the AIC values for each model on the aggregated data of all participants. The results of this analysis are fully consistent with the results based on BIC reported in the main text.

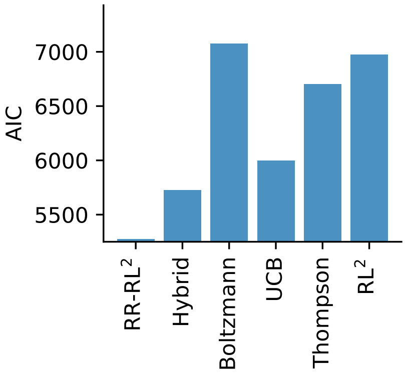
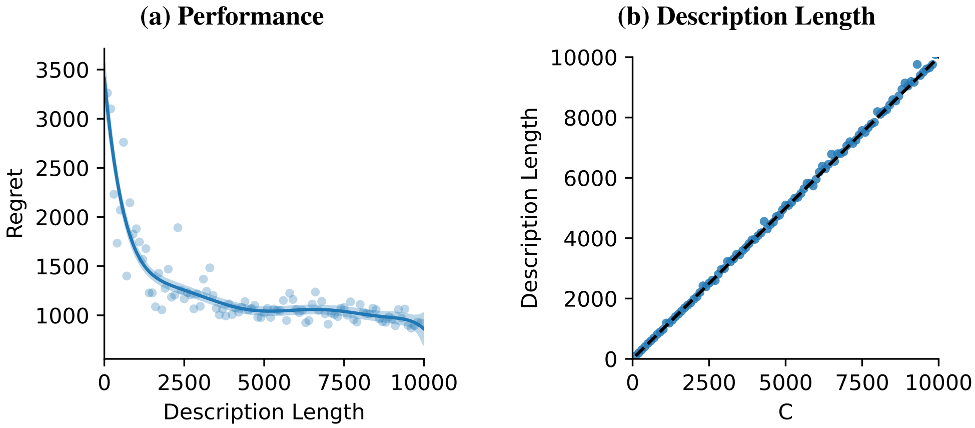


Figure C2: Akaike information criterion (AIC) values for the two-armed bandit data from Gershman [2018]. Lower values correspond to a better fit to human behavior.

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.



(a) Performance

(b) Description Length

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.

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

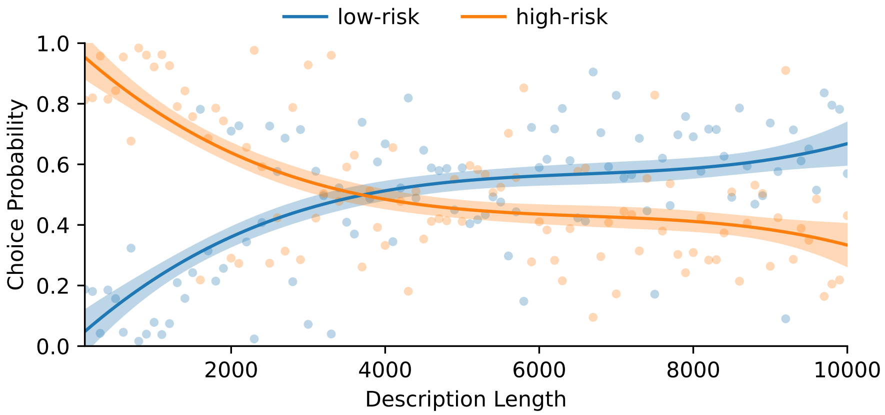


Figure D2: Probability of selecting low- and high-risks arms in the Iowa Gambling Task for all description lengths. Mean choice probabilities are illustrated as solid lines obtained by fitting a Bayesian polynomial regression model to the underlying data. Shaded contours represent the standard deviation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

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.

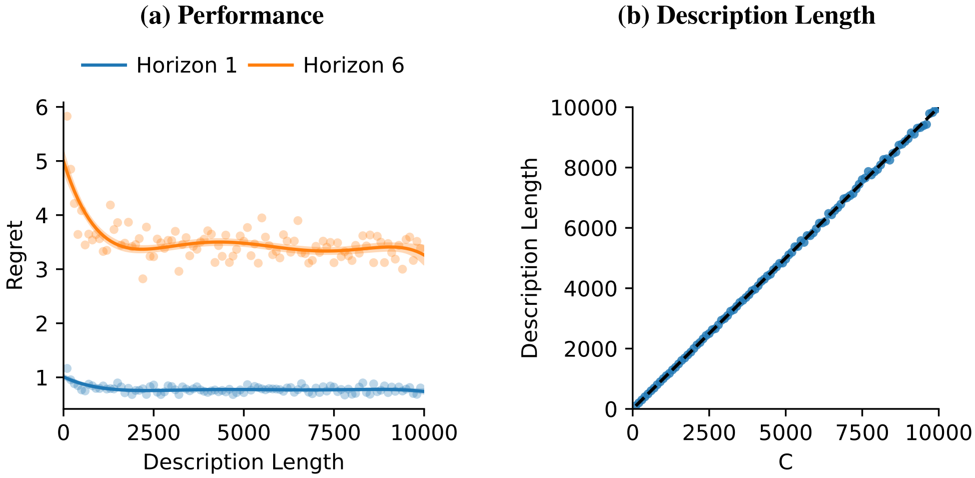


Figure E1: Performance and description lengths in the horizon task. (a) Regrets for models with varying description lengths. The solid line shows the mean prediction of a Bayesian polynomial regression model. Shaded contours represent the standard deviation of the mean. (b) Targeted description lengths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