Zero-shot compositional reinforcement learning in humans

Akshay K. Jagadish[1,+], **Marcel Binz**[1], **Tankred Saanum**[1], **Jane X. Wang**[2], **and Eric Schulz**[1,*]

[1]MPRG Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics
[2]Google Deepmind
[+]akshay.jagadish@tue.mpg.de
[*]eric.schulz@tue.mpg.de

ABSTRACT

People can easily evoke previously learned concepts, compose them, and apply the result to solve novel tasks on the first attempt. The aim of this paper is to improve our understanding of how people make such zero-shot compositional inferences in a reinforcement learning setting. To achieve this, we introduce an experimental paradigm where people learn two latent reward functions and need to compose them correctly to solve a novel task. We find that people have the capability to engage in zero-shot compositional reinforcement learning but deviate systematically from optimality. However, their mistakes are structured and can be explained by their performance in the sub-tasks leading up to the composition. Through extensive model-based analyses, we found that a meta-learned neural network model that accounts for limited computational resources best captures participants' behaviour. Moreover, the amount of computational resources this model identified reliably quantifies how good individual participants are at zero-shot compositional reinforcement learning. Taken together, our work takes a considerable step towards studying compositional reasoning in agents – both natural and artificial – with limited computational resources.

Introduction

People have an impressive ability to learn from sparse data[1]. We can acquire a new word from only one encounter[2]. We can achieve near-perfect classification rates from only one labelled observation[3]. We can even ask questions such as "how likely is it that a newly invented machine could transform a man into a vase?"[4], even though we are unlikely to ever encounter such a machine. Many researchers have proposed that the ability to generalize from sparse data is a hallmark of human intelligence[5,6].

What are the mechanisms that underlie this ability? One mechanism that enables strong generalizations is compositionality[5,7–10], which is the idea that complex entities can be constructed through the combination of primitive elements. People are generally considered to excel at reasoning compositionally[11]. They can, for example, combine parts of objects into novel objects[7,12] or compose previously learned actions to explore in novel contexts[8,13]. It has thus been argued that compositionality equips us with the ability to *"make infinite use of finite means"*[14,15], allowing us to generalize to novel situations by reusing and combining past experiences[13,16,17].

Empirical studies have demonstrated that people have an inherent predisposition towards compositional patterns[7,18–23]. For example, utilizing the function learning paradigm, which involves the learning, completion, and prediction of functional patterns, Schulz and colleagues[18,19] have demonstrated that humans find it easier to learn about compositional than non-compositional patterns. Furthermore, they showed that humans exhibit superior abilities to complete and predict compositional functions, as well as an enhanced capacity for remembering such functions[20,21]. These findings extend beyond function learning to other domains such as spatial structure learning[23,24], concept learning[7,25], shape perception[26], and auditory sequence learning[27]. Taken together, there is strong evidence for the presence of compositional inductive biases in humans.

While the preference for compositional patterns has received significant attention, how people compose two already learned functions and act on them in a zero-shot manner remains less well-understood. We attempt to close this gap in the present paper by studying human compositional reasoning in a reinforcement learning setting. More specifically, we are interested in how people perform zero-shot compositional inferences on learned latent reward functions. To study this question, we propose a novel experimental paradigm in which people interact with a sequence of three structured multi-armed bandit tasks[20,28–30] as illustrated in Fig. 1. The rewards for the first two sub-tasks are sampled from differently structured functions. They are followed by a third sub-task in which the rewards are set to a composition of the previously encountered functions. The structure of our task induces a learning curriculum that allows participants to solve the final sub-task in a zero-shot manner – assuming that they can reason compositionally.

In two experimental studies, we find that people can make such zero-shot compositional inferences in a reinforcement

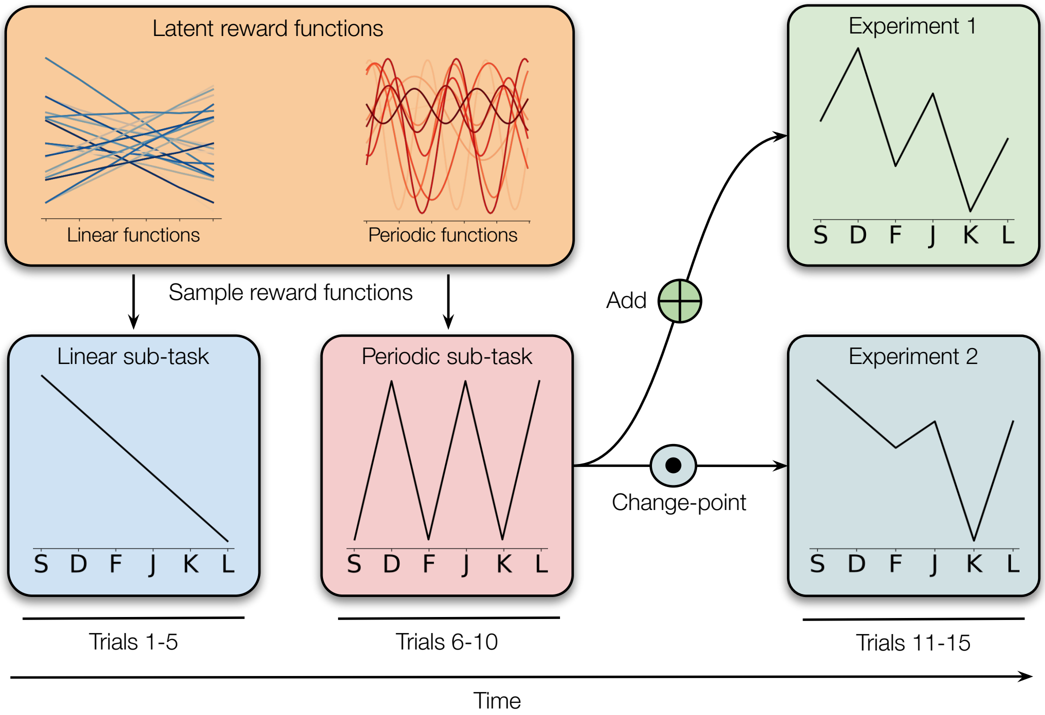

Figure 1. Overview of our experimental paradigm. Each task consists of three multi-armed bandit sub-tasks, with participants performing five trials per sub-task. Each multi-armed bandit task has six arms, each corresponding to different letters on the keyboard (starting from S to L). Rewards for the arms in our bandit tasks follow a latent function that is dependent on the spatial position of the arms. The reward functions for the first two sub-tasks are sampled from either the linear or the periodic family. The rewards for the final sub-task are constructed by composing the reward functions sampled in the two earlier sub-tasks. We conducted experiments with two different composition rules: an additive rule (experiment 1) and a change-point rule (experiment 2). In experiment 1, rewards for the final sub-task are constructed by performing an element-wise summation of sampled rewards from the first two sub-tasks. Whereas in experiment 2, they are constructed by combining segments of the sampled rewards such that the rewards change from being from one family to another after a certain number of options. Specifically, rewards in the first segment (i.e. options S-D-F) come from one family – either a linear or periodic family – and the alternative family in the second segment (i.e. options J-K-L). Note that the order in which these reward segments are combined in the change-point rule is randomized. If participants learn the latent reward functions in the first two sub-tasks and apply the compositional rule correctly, they can – in principle – perform zero-shot compositional inference, meaning that they choose the optimal arm (arm D in the example shown for experiment 1 and arm S in the example shown for experiment 2) on the first trial of the last sub-task.

learning setting. However, our analyses also indicate that their behaviour deviates systematically from a fully-normative account. Extensive model-based analyses furthermore reveal that human compositional reasoning is overall best explained by a resource-rational account[31,32]. Taken together, our results suggest that people can make zero-shot compositional inferences but that their performance is constrained by cognitive demands.

Results

To investigate compositional reinforcement learning in humans, we developed a novel multi-armed bandit paradigm based on previous works[20,29,30] as illustrated in Fig. 1. Each task consists of three multi-armed bandit sub-tasks in which rewards follow a latent function that is dependent on the spatial position of the arms. The reward functions for the first two sub-tasks are sampled from either the linear or the periodic family of functions. In the final sub-task, reward functions are constructed by composing the reward functions encountered in the two earlier sub-tasks. We conducted two experiments with different composition rules: an additive rule and a change-point rule. Our task induces a learning curriculum, which enables us to probe whether people are able to reason compositionally in a reinforcement learning setting. In particular, we expect people to solve the final sub-task in a zero-shot manner, selecting the best option on the first trial. To have a comparison, we also consider a condition without a curriculum. In this non-curriculum condition, people do not interact with the first two sub-tasks and instead directly observe the composite function from the final sub-task. We set the length of each sub-task to five, leading to 15 trials per task in the curriculum and five trials per task in the non-curriculum condition. Note that the number of trials per sub-task is less than the number of available options, thereby preventing participants from exhaustively trying out all options and forcing them to generalize based on the underlying function.

Experiment 1: Additive rule

We conducted an online behavioural study, on the Prolific platform, following the structure of the just outlined task to test the underlying mechanisms behind how people compose. Participants played a game under a cover story that they were interacting with slot machines produced by two manufacturers (Blue Lagoon and Green Geeks). They were told that all slot machines from the same manufacturer behaved similarly. However, participants were not told which manufacturer each slot machine belonged to, but had to figure this out through trial and error. In the curriculum condition, participants played with a slot machine from each manufacturer before playing a compositional slot machine that combined the two. In the non-curriculum condition, participants only played with the compositional slot machine. The study involved 20 tasks per participant, leading to 300 trials in total for the curriculum condition and 100 for the non-curriculum condition. We provide further details about the experiment and participants in the Materials and Methods section.

Behavioural analysis

First, we wanted to establish that people successfully composed reward functions in our task. For the corresponding analyses, we considered two behavioural measures: regrets and the probability of making optimal choices. Fig. 2(a) shows the mean regret of participants in the compositional sub-task. The regret is computed by taking the difference between the highest reward in the given sub-task and the reward for the action selected by the participant. Participants are said to have successfully performed zero-shot compositional inference if they choose the optimal arm on the first trial of the compositional sub-task (or have a regret measure of 0 on the first trial). We see from the regrets that people performed better than chance right from the outset for the curriculum condition (Mean $(M) = 2.163$, Standard Error $(SE) = 0.116$; t[1] $= -19.57, p < .001$) whereas they start at chance-level for the non-curriculum condition ($M = 4.055$, $SE = 0.059$). To further quantify the effects of zero-shot compositional inference, we performed a mixed-effects linear regression on the regrets in the final sub-task with trials, conditions, and their interaction as fixed effects (and with random slopes and intercepts per participant for all of these factors). This analysis revealed that participants in the curriculum condition had a significantly lower regret on the first trial of the final sub-task than participants in the non-curriculum condition ($\hat{\beta} = -1.18 \pm 0.115; z = -10.24, p < .001$). In addition, a comparison of the probability for making an optimal choice on the first trial between curriculum and non-curriculum conditions, shown in Fig. 2(b), confirmed that people in the curriculum condition ($M = 0.382$, $SE = 0.02$) made optimal choices more frequently than in the non-curriculum condition ($M = 0.192$, $SE = 0.07$; $t = -9.242, p < .001$). Regret performance on the first trial of the curriculum condition was even better than performance on the last trial of the non-curriculum condition ($M = 2.40$, $SE = 0.13$; $t = 2.72, p < 0.01$), suggesting that learning within the last sub-task cannot match the performance boost gained from compositional inference.

While people were able to compose in a zero-shot manner, they did not do so perfectly. Their initial regrets in the final sub-task ($M = 2.163$, $SE = 0.116$; $t = 34.60, p < .001$) deviated significantly from ideal compositional reasoning. Further evidence of people's suboptimality comes from the observation that they continued learning during the final sub-task in the curriculum condition (which would not be needed if they were to engage in perfect zero-shot compositional inference). To quantify this effect, we fitted a mixed-effects linear regression model using per-trial regret in the last sub-tasks as the dependent variable, and the corresponding trial number as both fixed effects and random effects over participants. The results of this model showed a significant fixed effect of trial number ($\hat{\beta} = -0.32 \pm 0.02; z = -13.88, p < .001$) onto regret, confirming that the performance of participants improved with additional interactions. The observed improvement in the curriculum condition ($\hat{\beta} = -0.21 \pm 0.02; z = -12.26, p < .001$) was generally weaker than that in the non-curriculum condition ($\hat{\beta} = -0.42 \pm 0.02; z = -21.80, p < .001$).

We also inspected the marginal action distribution of participants on the first trial of the final sub-task shown in Fig. 2(c). We see that the mode of the participants' action distribution matches the optimal choice, but that human behaviour also systematically deviates from optimal behaviour. Particularly, one interesting feature is that people seem to pick corner arms – especially the left-most one – frequently. This could reflect a bias that has been observed in other studies of people exploring different options starting from left to right[29].

To better understand the mistakes that people make during compositional reasoning, we looked at how participants' performance in the first two sub-tasks can explain their behaviour on the final compositional sub-task. We first classified choices on the first trial of the compositional sub-task into four categories: first, picking the optimal arm as predicted by compositional inference; second, a non-optimal corner arms category which includes trials where people picked the corner

[1] t-values are reported from a non-parametric independent two-sample t-test (two-tailed) using 1000 random permutations

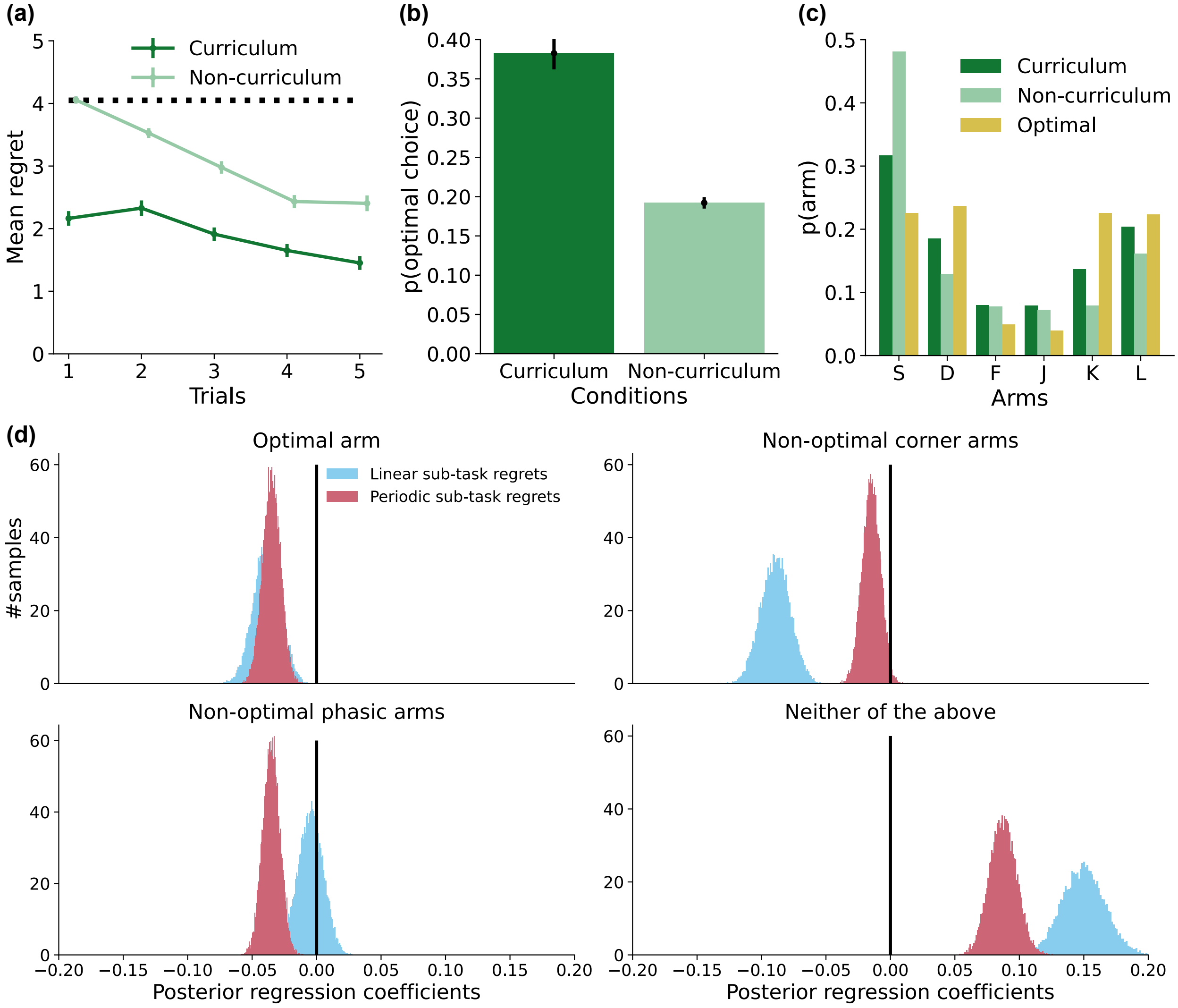

Figure 2. Behavioural results of experiment 1. (a) Mean regrets for participants in the final sub-task for the two conditions: curriculum and non-curriculum. The dotted lines in black indicate mean regret from a random policy. (b) Probability of participants making the optimal choice on the first trial of the final sub-task for the curriculum and non-curriculum condition. Error bars in (a) and (b) represent standard errors computed over participants. (c) Marginal distribution of choices on the first trial of the final sub-task for the two conditions. The bar in gold shows the marginal distribution of choices for the optimal policy. (d) Explaining choices of participants in the last sub-task based on their task performance in the first two sub-tasks. We classified the choices on the first trial into four categories: optimal arm (top-left), non-optimal corner arms (top-right), non-optimal phasic arm (bottom-left), and neither of the above (bottom-right). Then, we fit a Bayesian logistic regression model from the total regrets (summed over all trials) in the first two sub-tasks onto each of these categories. The sub-plots show the histogram of the posterior regression coefficients of linear and periodic sub-tasks for all four choice categories.

arms despite them not being the optimal choice; third, a non-optimal phasic arms category which includes trials where arms belonging to the same phase as the periodic sub-task were picked even when it was not the optimal choice; and fourth, a category which includes all trials where choices did not fall into any of the three categories mentioned above. We then fitted a separate Bayesian logistic regression model in PyMC3 from the total regrets (summed over all trials) in the first two sub-tasks onto each of these four choice categories coded as a binary variable. Fig. 2(d) visualized the posterior regression coefficients from these fitted models with each category shown in a separate sub-plot. When the dependent variable was picking the optimal choice, the posterior regression coefficients were negative with similar means for both linear ($M = -0.038$, $SE = 7.319e-05$) and periodic ($M = -0.035$, $SE = 4.925e-05$; $t = -40.728, p < 0.001$) sub-tasks. This suggests that participants pick the optimal arm when they learn both the sub-tasks well. The coefficients for regrets of the linear sub-tasks were more negative ($M = -0.089$, $SE = 8.107e-05$) than that of periodic regrets ($M = -0.015, SE = 5.040e-05; t = -779.036, p < .001$) when the dependent variable was non-optimal corner arms. This result suggests that people tend to pick non-optimal corners arms when they learned linear functions better than periodic functions. When the dependent variable was non-optimal phasic arms, the posterior regression coefficients for regrets from the periodic sub-task were lower ($M = -0.0352, SE = 0.477e-04$) than that of the linear sub-task ($M = -0.0048$, $SE = 7.0762e-05$; $t = 356.072, p < 0.001$). This result indicates that people pick one of the phasic arms from the periodic sub-task on the first trial of the compositional sub-task when they perform better in the periodic sub-task. Lastly, the regression coefficients were positive for both the regrets from the linear sub-task ($M = 0.1498$, $SE = 1.167e-05$; $t = 450.07, p < 0.001$) and those from the periodic sub-task ($M = 0.0874$, $SE = 7.502e-05$) when the dependent variable was neither of the categories above. This result suggests that people pick neither the optimal, the corner nor the phasic arm when they have not learned the underlying functions in either of the two sub-tasks well.

Taken together, behavioural results from experiment 1 suggest that people can compose in a zero-shot fashion but are not perfect. However, their mistakes are highly structured and can be predicted based on how well they have learned the different components of the first two sub-tasks.

Model-based analysis

In our compositional bandit task, people can – in principle – perform near-perfect if they manage to compose. However, the behavioural analysis above revealed that people (despite generally managing to compose) systematically deviated from optimal behaviour. To get a better understanding of these deviations and the cognitive processes behind them, we investigated people's behaviour using computational models.

We considered six different computational models for explaining participants' choices in our task. Four of these are Bayesian models that vary along two dimensions: first, whether or not they can generalize learned values from one option to the other, and second, whether or not they can compose the learned values from the first two sub-task to reason on the final sub-task. The Bayesian models include a Bayesian mean-tracker (BMT)[33] which is a model that does not learn about the underlying functional structure but instead updates its beliefs about rewards for each option independently, as well as a model that learns functions by generalizing across options within a sub-task based on the idea of Gaussian Process regression (GPR)[18,34,35]. For each of these two models, we considered one variant that cannot compose and instead learns separate reward functions for each sub-task, and another one that initializes its predictions in the final sub-task to the composition of the learned means from the first two sub-tasks.

In addition, we also considered two recurrent neural network models that were trained via meta-reinforcement learning[36,37]. Unlike the Bayesian models from above, these models learn inductive biases about latent reward functions via trial-and-error, without requiring an explicit specification of priors. The first of these models is RL^2 – a model that is known to approximate the Bayes-optimal policy for the distribution of tasks it was trained on[38,39], which thereby allows us to test whether people compose optimally. The second is a resource-rational extension of RL^2 referred to as RR-RL^2[40]. The particular resource constraint considered by RR-RL^2 is the description length of the meta-learned recurrent neural network, which is defined as the number of bits required to store its parameters. RR-RL^2 captures the hypothesis that people attempt to achieve optimal performance but that they are subject to the constraint of relying on an algorithm with limited computational complexity. We fitted RR-RL^2's description length on a participant-by-participant basis reflecting the assumption that different participants use different amounts of computational resources.

We simulated all of these models on our compositional bandit task and measured their performance in the final sub-task. We found that performance on the first trial was near-optimal for the four models that can compose (the compositional BMT and GPR, as well as the two meta-learning agents), indicating that they can re-use the earlier learned functions to compose new functions in a zero-shot manner; for detailed visualizations, see Supplementary Information (SI).

To obtain a quantitative measure of the goodness-of-fit of models to human choices, we conducted a Bayesian model comparison of all previously outlined models. We measured the fit to human choices based on two metrics: posterior model frequency and exceedance probability[41]. The posterior model frequency measures how often a model offers the best explanation in the population, while the exceedance probability measures how likely it is that a given model is the most frequent explanation. Further details about this model comparison procedure can be found in the Materials and Methods section.

This model comparison revealed that RR-RL^2 captures how people behave on the first trial of the compositional sub-task the best according to both metrics, with exceedance probability amounting to 0.99, while its posterior model frequency was 0.704 ± 0.002. The compositional BMT is the second-best model with $4.33e-09$ and 0.288 ± 0.002 on exceedance and posterior model frequency respectively. Interestingly, we found that the two models that performed best in our model simulations (compositional GPR and RL^2) did not predict human behaviour well. Taken together, these results support the hypothesis that people do not compose in a fully optimal way, but that their ability to reason compositionally is driven by principles of resource rationality.

To further support the model comparison results, we simulated behaviour from the two best-fitting models and compared

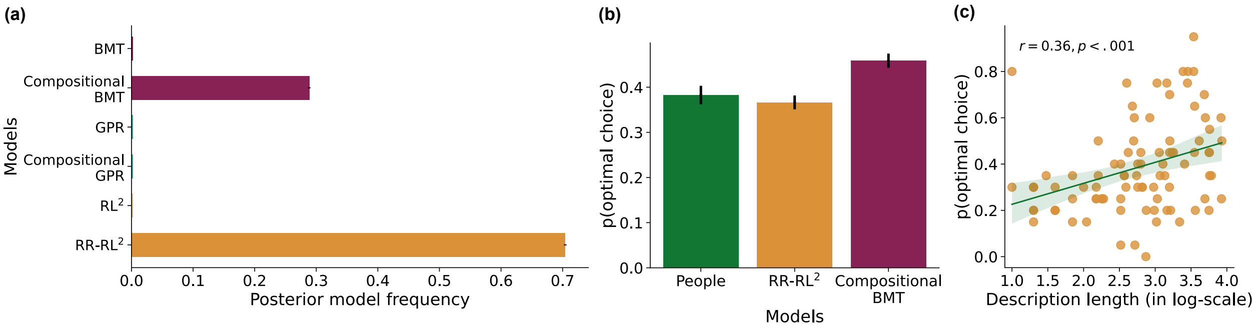

Figure 3. Modelling results of experiment 1. (a) The posterior model frequency of participant choices on the first trial of the last sub-task. (b) Comparison of the probability of making the optimal choice on the first trial of the last sub-task between people and simulations from the best-fitting models: RR-RL^2 and compositional BMT. (c) Correlation between fitted description length of RR-RL^2 (plotted in log-scale) and the probability of making an optimal choice on the first trial of the last sub-task. The fitted regression line is shown in green, with the shaded portion showing the 95% confidence interval.

them against human behaviour. With regards to the probability of making the optimal choice on the first trial, we found that simulations from RR-RL^2 ($M = 0.366$, $SE = 0.0154$) matched human behaviour ($M = 0.382$, $SE = 0.021$) whereas the compositional BMT differed significantly ($M = 0.459$, $SE = 0.016$; $t = -2.938, p < .01$).

Next, we examined whether the fitted description lengths of RR-RL^2 could capture task performance. To do this, we correlated the probability of humans making the optimal choice on the first trial against the fitted description lengths for each participant. We found that description length correlated significantly with optimality ($r = 0.359$, $p < .001$) as illustrated in Fig. 3(c). Likewise, we also observed a significant negative correlation ($r = -0.36, p < .001$) between fitted description lengths and mean regrets on the first trial.

We then analysed whether fitted description lengths can be used to explain the types of choices participants make. For this, we grouped participants based on their fitted description lengths into two groups. In the first group, we included participants whose fitted description lengths were in the range of 1000 to 10000 ($N = 39$), and in the second group, we considered those whose fitted description lengths were in the range of 10 to 100 ($N = 17$). We then compared the probability of making the optimal choice on the first trial of the final sub-task between the two groups. We found that participants in the first group performed near-perfect composition, whereas the choices from participants in the second group were far away from optimality (for further details, see SI). Thus, fitted description lengths can be used to cluster participants into those who can perform near-optimal zero-shot compositional inference and those who cannot.

Lastly, we looked at how regrets on the first two sub-tasks influence behaviour on the first trial in RR-RL^2, just like how we did it in people. We found that the posterior regression coefficients of the model match human behaviour qualitatively with slight deviations. The model picks the non-optimal corner arms and non-optimal phasic arms when it learns one sub-task better than the other (with performance in periodic sub-task having a greater influence on the linear in both cases) and follows a completely different strategy from the ones above when it does not learn both the sub-tasks well. An interesting deviation was that, unlike humans who make more optimal choices when they learn both sub-tasks equally well, the model does so when they learn the linear sub-task better than the periodic sub-task (for detailed results and visualizations, see SI).

Taken together, results from behavioural and model-based analyses suggest that people can perform zero-shot compositional inference but still deviate systematically from optimal behaviour. Their choices are best explained by a meta-reinforcement learning model (RR-RL^2) that learns a solution with limited computational resources. Furthermore, we find the simulated behaviour from RR-RL^2 matches human behaviour well and that description length – the parameter that controls computational resources of RR-RL^2 – correlates with the probability of making an optimal choice on the first trial of the compositional sub-task.[2]

Experiment 2: Change-point rule

Next, we wanted to verify that the results obtained in the previous section transfer to another compositional rule. We, therefore, conducted a second experiment using a change-point rule. The experimental procedure followed the same structure as the

[2] Note that in the main text, we have focused on model-comparison results for the first trial of the compositional sub-task as we are mostly interested in zero-shot compositional inference. However, we also evaluated our models over all trials of the compositional sub-task. The results of these analyses are summarised in the SI.

additive rule but with participants now being tested on a slot machine whose rewards were composed based on the change-point rule (see Fig. 1 for an example).

Behavioural analysis

Like in the additive rule experiment, we see that people perform much better than chance right from the outset for the curriculum condition ($M = 1.937$, $SE = 0.081$; $t = -15.105, p < .001$) while starting at chance-level for the non-curriculum condition ($M = 3.096$, $SE = 0.060$) as shown in Fig. 4(a). We also performed a mixed-effects linear regression analysis as we did in the first experiment which confirmed that participants in the curriculum condition had a significantly lower regret on the first trial of the final sub-task than participants in the non-curriculum condition ($\hat{\beta} = -1.18 \pm 0.115; z = -10.24, p < .001$). When looking at the probability of picking the optimal arm in Fig. 4(b), we also find that people make better choices in the curriculum condition ($M = 0.337$, $SE = 0.016$) than in the non-curriculum condition ($M = 0.168$, $SE = 0.008$, $t = -9.347, p < .001$). The performance in the curriculum condition is better than in the non-curriculum condition for all trials, which was also the case in the additive rule. Thus, similar to the additive rule experiment, people are able to perform approximate zero-shot compositional inferences.

Even though people were able to compose in a zero-shot manner, they were again not flawless. Their initial regrets in the final sub-task ($M = 1.937$, $SE = 0.081$; $t = 38.25$, $p < .001$) deviated significantly from ideal compositional reasoning, thereby corroborating our results from the previous experiment. In addition, people's suboptimality was underlined by the persistent presence of learning effects in the curriculum condition ($\hat{\beta} = -0.216 \pm 0.013; z = -16.112 p < .001$).

We also inspected participants' marginal action distribution on the final sub-tasks first trial in Fig. 4(c). We see that the mode of the participants' action distribution lies at the optimal choice. However, human behaviour systematically deviates from optimal behaviour as people pick sub-optimal options frequently. People also tend to pick the corner options – especially the left-most option – quite frequently as they did in the additive rule.

Finally, we repeated the regret analysis we did for the additive rule to better understand which kind of mistakes people make. The results of this analysis are summarized in Fig. 4(d). We see that posterior regression coefficients for regrets of both linear ($M = -0.0423, SE = 7.518e-05$) and periodic ($M = -0.0464, SE = 5.038e-05$) sub-tasks are negative and have overlapping distributions in cases where people picked the optimal option ($t = 45.874, p < .001$). This suggests that learning both linear and periodic sub-tasks equally well predicts good performance on the first trial. When people pick non-optimal corner arms, their performance in the linear sub-task ($M = -0.0607, SE = 7.663e-05$) seems to be driving their behaviour more than their performance in the periodic sub-task ($M = -0.0305, SE = 5.0042e-05$). However, on the contrary, picking the non-optimal phasic arm is not driven strongly by periodic sub-task performance with regression coefficients for both linear ($M = -0.0259, SE = 7.0662e-05$) and periodic ($M = -0.0203, SE = 4.672e-05$) sub-tasks overlapping ($t = -66.174, p < 0.001$). Lastly, the posterior regression coefficients are distributed on the positive axis for both linear ($M = 0.114, SE = 9.796e-05$) and periodic ($M = 0.0747, SE = 6.368e-05$) regrets when predicting non-optimal choice belonging to neither to corner or phasic arms.

Model-based analysis

The results for model-based analyses with the change-point rule mirrored that of the additive rule. We find again that RR-RL^2 captures best how people compose according to both metrics with exceedance probability amounting to 0.99, while its posterior model frequency was $0.741 \pm 1.729e-03$. The compositional BMT is the second-best model with an exceedance probability of close to 0 and a posterior model frequency of $0.253 \pm 1.702e-03$ as visualized in Fig. 5(a). These results thus show again that people do not compose in a fully optimal way, but that their ability to reason compositionally is instead impeded by computational constraints.

We simulated behaviour from the two best-fitting models with their parameters fitted to participant behaviour. Looking at the probability of selecting the optimal choice, we found that simulations from RR-RL^2 ($M = 0.2922, SE = 0.0132$) matched human behaviour ($M = 0.3372, SE = 0.0158$) more closely than the compositional BMT ($M = 0.2862, SE = 0.0069$; $t = 2.9560, p < .01$) as shown in Fig. 5(b).

We found that fitted description lengths correlated significantly with the probability of participants making an optimal choice for the change-point rule as well, showing a correlation coefficient of $r = 0.487$ ($p < 0.001$) as shown in Fig. 5 (c). The result also holds when we use regrets as a performance measure ($r = -0.448, p < 0.001$).

Following our earlier analysis, we grouped the participants based on their fitted description lengths into two groups with the first group including participants with fitted description lengths between 1000 to 10000 ($N = 58$) and the second group including participants with fitted description lengths between 10 to 100 ($N = 25$). We compared zero-shot compositional inference between the two groups (for detailed analysis, see SI) and found again that participants in the first group performed near-perfect zero-shot compositional inferences, whereas the choices from participants in the second group were far away from optimality.

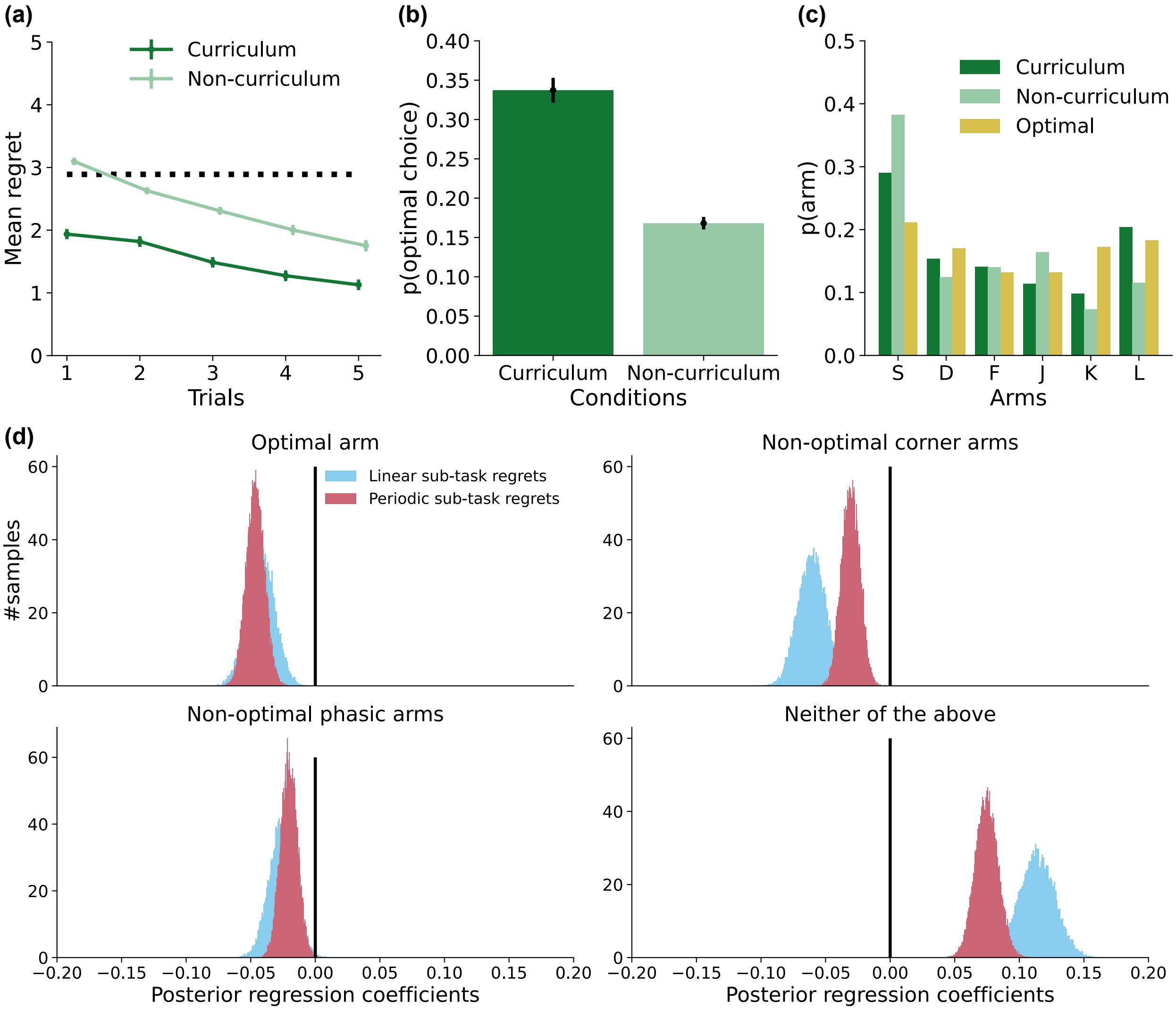

Figure 4. Behavioural results of experiment 2. (a) Mean regrets for participants in the final sub-task for the two conditions: curriculum and non-curriculum. The dotted lines in black indicate mean regret from a random policy. (b) Probability of participants making the optimal choice on the first trial of the final sub-task for the curriculum and non-curriculum condition. Error bars in (a) and (b) represent standard errors computed over participants. (c) Marginal distribution of choices of participants on the first trial of the final sub-task for the two conditions. The bar in gold shows the marginal distribution of choices for the optimal policy. (d) Explaining choices of participants in the last sub-task based on their task performance in the first two sub-tasks. The choices on the first trial were classified into four categories: optimal arm (top-left), non-optimal corner arms (top-right), non-optimal phasic arm (bottom-left), and neither of the above (bottom-right). Then, we fit a Bayesian logistic regression model from the total regrets (summed over all trials) in the linear and periodic sub-tasks onto each of these categories. The sub-plots show the histogram of the posterior regression coefficients of linear and periodic sub-tasks for all four choice categories.

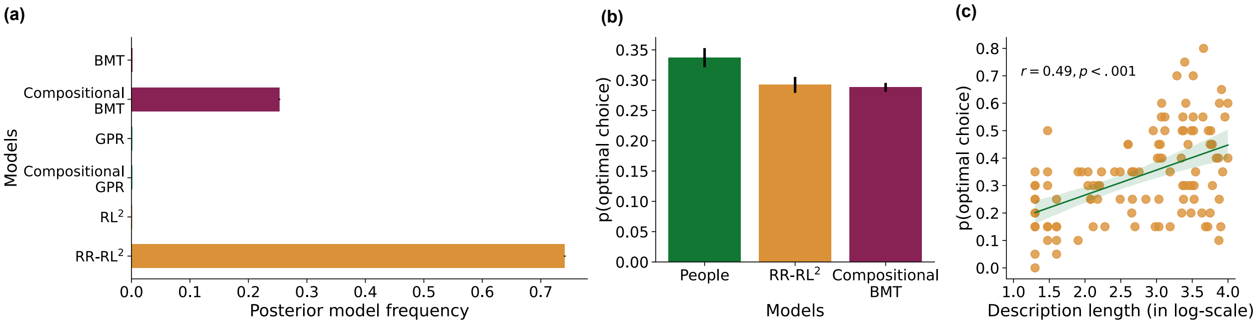

Figure 5. Modelling results of experiment 2. (a) The posterior model frequency of participant choices on the first trial of the last sub-task. (b) Comparison of the probability of making the optimal choice on the first trial of the last sub-task between people and simulations from the best-fitting models: RR-RL^2 and compositional BMT. (c) Correlation between fitted description length of RR-RL^2 (plotted in log-scale) and the probability of making an optimal choice on the first trial of the last sub-task. The fitted regression line is shown in green, with the shaded portion showing the 95% confidence interval.

Finally, we inspected how performance on the first two sub-tasks influenced behaviour on the first trial of the compositional sub-task in RR-RL^2. We found that overall the posterior regression coefficients of the model's behaviour in linear and periodic sub-tasks match human behaviour quite well. They pick the non-optimal corner arms and non-optimal phasic arms when they learn one sub-task better than the other and they follow a completely different strategy from the ones above when they do not learn both the sub-tasks well. However, as in experiment 1 but to a smaller extent, the model makes optimal choices when they learn the linear sub-task better than the periodic sub-task. The results of these analyses are summarised in the SI.

Taken together, these results mirror those we had for the additive rule with similar factors affecting human behaviour in the compositional sub-task. This suggests that people's ability to do compositional inference is robust with regard to the specific way in which functions are composed.

Discussion

Compositionality is at the core of people's ability to generalize from sparse data. It has even been argued to be an essential component of intelligence more generally[5,7,11]. However, how people use this ability to make decisions is less well understood. To address this question, we have proposed a novel experimental paradigm where people first need to learn about latent reward functions in two sub-tasks to be able to pick the most rewarding option on the first trial of the third sub-task. We found that people indeed perform this kind of zero-shot compositional inference, but they deviate systematically from ideal behaviour. Even so, their mistakes were not random but instead highly structured. Extensive model-based analyses revealed that RR-RL^2 – a meta-learned neural network model that accounts for limited computational resources – captures participants' behaviour the best. Mistakes made by this model were also systematic and predicted by similar factors that predicted human suboptimal choices. This result indicates that people seem to follow resource-rational principles when making compositional inferences, thereby expanding on earlier results from other cognitive domains such as decision-making[42], planning[43,44], and problem-solving[45].

Relation to empirical works in compositional reinforcement learning

Previous work has investigated how people structure prior knowledge and compose specific components of this learned knowledge to generalize efficiently in reinforcement learning tasks. Xia and Collins[13] have used the options framework from hierarchical reinforcement learning to show that humans can learn hierarchical options and do so such that the temporal ordering of the learned options remains intact. They further showed that people, akin to their model, can compose the learned options to explore novel contexts, thereby speeding up learning. Looking into how learned knowledge guides generalisation, Franklin and Frank[8] showed that human learners decompose learned task structures into distinct components such as rewards and state transitions. They devised a meta-learning agent[46] that trades off re-using these components jointly or compositionally and showed that, similar to this agent, humans too act adaptively on these components depending on the statistics of the task environment. In comparison to our work, these previous studies focused on how people generalize in a sample-efficient way to novel tasks and not on zero-shot compositional inference. Therefore, our work complements these earlier investigations by addressing a distinct dimension of compositional reasoning. It is worth noting that adapting the proposed models to our experimental paradigm is not straightforward. However, our compositional GPR shares a similar flavour to the hierarchical models proposed by Xia and Collins, indicating potential conceptual connections between different approaches to compositional reasoning.

Model complexity and compositional inference

The relationship between description length and human performance has received considerable attention in previous research[20,23–25]. However, researchers have typically investigated how the description length of the *task* influences performance. For instance, Amalric and colleagues[24] asked participants to predict and repeat sequences displayed on a clock-like display, varying the complexity of the sequences by changing the length of the generating program. They found a correlation between the difficulty of predicting (and repeating) a given sequence and its complexity. In contrast, we investigated how the description length of *strategies* applied by individual participants influenced their performance. We found that different participants apply strategies with different description lengths, implying that they use varying amounts of cognitive resources to perform the task. Thus, task performance is influenced not only by external factors such as task difficulty but also by internal factors such as individual differences in cognitive resources.

Neural networks and compositionality

In contrast to the prevailing notion that neural network models struggle with compositional reasoning[47,48], we found that the model that captured human behaviour best in our experiment was a neural network model, suggesting that these models are not inherently unable to reason compositionally. Instead, it matters how they are set up and how they are trained. This result is supported by other recent studies demonstrating that neural network models can be good models of human compositionally[49–53]. For example, Lake[49] demonstrated that meta-learning can be used to train sequence-to-sequence networks that generalize compositionally in human-like ways on the SCAN data set[47], while Kumar and colleagues[50] found that augmenting meta-reinforcement learning agents with an auxiliary objective to reproduce task descriptions aligns them with human behaviour in a setting that requires reasoning about compositionally-generated patterns. Finally, Dekker and colleagues[51] designed neural network architecture with an inductive bias for compositional reasoning using a Hebbian gating process, and demonstrated that the resulting model learns composable functions similar to how these functions are learned by people.

Limitations

While one might argue that the proposed compositional bandit task is too simplistic, we found that human behaviour systematically deviated from optimality, implying that the task complexity was appropriate for the investigated research question. Furthermore, our task has two main advantages compared to those previously used to study compositional reasoning. First, it is directly inspired by experiments used to study human learning in structured environments[29], allowing us to connect our findings to previous work on human cognition. The second advantage is its simplistic design. This simplicity allowed us to build computational models that solve the task near-optimally, picking the best option in a zero-shot fashion. Nevertheless, it might be interesting to develop more naturalistic compositional reasoning tasks in future work to test if our model-based predictions still hold. There are additional variants of our paradigm that could be considered. For example, one could test whether increasing the length of the first two sub-tasks causes people to make better compositional inferences. People's performance could furthermore be boosted by relying on a purely observational setting in which options in the first two sub-tasks are presented in a structured manner (for example from left to right).

We also note two shortcomings on the modelling side. First, we did not consider resource-rational Bayesian models in our model-based analyses, as building such models is not straightforward. In contrast to this, limited resources are easy to account for in the meta-learning setting[39] which is why we relied on such models instead. Second, our models assume each task to be independent of each other. This may be in contrast to humans who could show learning-to-learn effects across the entire experiment. For example, it might be the case that some participants need a few trials to apply the compositional rule correctly or need a few tasks to correctly learn the underlying functions. It might even be the case that some participants start doing compositional reasoning only after an *Aha!* moment[54]. Learning-to-learn effects could be incorporated into both the Bayesian and the meta-learned models. For the Bayesian models, one could build on the work of[29] who used a simple clustering algorithm to capture the learning-to-learn behaviour across tasks. For meta-learning agents, it would in principle be possible to train over samples of entire experiments (i.e., 20 successive tasks) where each experiment is sampled from a parameterized distribution of experiments. However, training such agents is challenging in practice especially since gradients need to be propagated over longer horizons.

Conclusion

We introduced a novel experimental paradigm and two complementary computational approaches for studying zero-shot compositional reinforcement learning in people. We showed that while people can perform zero-shot compositional inference in our task, their choices were better explained by a resource-constrained model than by optimal zero-shot compositional inference. Thus, our results provide a new perspective to the understanding of human compositional inference by considering the influence of cognitive resources. Taken together, our work takes a considerable step towards understanding compositional reinforcement learning in humans, symbolic, and sub-symbolic agents under computational constraints.

Materials and Methods

In this section, we provide details of the experimental methods and computational models used to analyse compositional reinforcement learning in humans. In the experimental methods subsection, we describe the task parameters, experimental design, participants, and ethics approval. In the computational methods subsection, we expand on the computational models and explain the methods used to fit these models to human behaviour along with the model comparison procedure.

Experimental methods

Participants

We recruited 200 participants (103 female, M_{age} = 28.90) through the Prolific platform for experiment 1. The study was approved by the local ethics committee. Participants were randomly assigned to the curriculum or non-curriculum condition. All participants had an approval rate of 95% or more, were fluent English speakers from the United States, and were 18 years of age or older. Participants were rewarded a base payment of £2 and a performance-dependent bonus payment up to £2.5. For experiment 2, we recruited 211 participants (96 females, M_{age} = 27.58) through the Prolific platform. The rest of the study parameters remained the same as the additive rule.

Task

Each task consisted of three multi-armed bandit sub-tasks in which rewards follow a function that is dependent on the spatial position of arms:

$$r_t = f(a_t) + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0, 0.1) \tag{1}$$

where t denotes the time-step, $a_t \in \{0, \ldots, 5\}$ the arm selected in time-step t and ε_t is an additive noise term. Reward functions f_{linear} and f_{periodic} for the first two sub-tasks are sampled from either the linear or the periodic family as shown below:

$$f_{\text{linear}}(a_t) = \left(\frac{2a_t}{5} - 1\right) w + b + \zeta \qquad w \sim \mathcal{U}(-2.5, 2.5), b \sim \mathcal{U}(2.5, 7.5) \tag{2}$$

$$f_{\text{periodic}}(a_t) = A\left|\sin\left(0.5\pi\left(a_t - \phi\right)\right)\right| + b + \zeta \qquad A \sim \mathcal{U}(0, 7.5), \phi \in \{0, 1\}, b \sim \mathcal{U}\left(0, \frac{A}{1.4}\right) \tag{3}$$

where $\mathcal{U}(a, b)$ is a uniform distribution on the interval [a,b] and $\zeta \sim \mathcal{N}(0, 0.2)$ is an additive noise term. The parameters were chosen after several rounds of piloting to make it easy for participants to perform the task well on average. For example, we found that excluding linear functions with very low slopes and periodic functions with very small amplitudes helped them learn the periodic and linear sub-tasks more easily and hence, improved their performance on the task overall.

Reward functions in the final sub-task were constructed by composing the reward functions encountered in the two earlier sub-tasks. We considered two different composition rules, an additive rule and a change-point rule:

$$f_{\text{additive}}(a_t) = f_{\text{linear}}(a_t) + f_{\text{periodic}}(a_t) \tag{4}$$

$$f_{\text{change-point}}(a_t) = \begin{cases} f_{\text{linear}}(a_t) & \text{if } a_t \in \{0, 1, 2\} \\ f_{\text{periodic}}(a_t) & \text{otherwise} \end{cases} \tag{5}$$

The order of composition in the change-point function was randomized. We set the length of each sub-task to 5 trials, leading to 15 overall trials per task. Note that the number of trials per sub-task was less than the number of available options. This prevents an agent from exhaustively trying out all options and forces it to generalize based on the underlying function.

Experimental design

We conducted an online behavioural study following the structure of the compositional bandit task outlined earlier to test the underlying mechanisms behind how people compose. Participants were told that they were gamblers visiting the fictional town of “Bandit City”. They visited multiple casinos (20 in total) in which they played different sets of slot machines. Each casino had two slot machines made by two different companies, called Blue Lagoon and Green Geeks, with their colour (included in the name) indicating the manufacturer. Participants were informed that all slot machines from the same company behaved similarly (i.e. rewards are sampled from the same underlying function, Equation 2 or 3), but were not told which reward function belonged to which company. They had to figure this out via trial and error during the experiment. However, they were shown two canonical examples from each reward function in the task instruction phase to get an idea of how samples from these reward functions could look like. In each casino, participants had five trials per slot machine, with the goal of winning as many coins as possible. In the curriculum condition, they would first interact with a slot machine from each of the two manufacturers. Following their interactions with the two slot machines, participants were tested on a new slot machine, which was a composition of the two previously played machines. Thus, in the curriculum condition, participants interacted with bandits for a total of 300 trials (breakup: 5 trials per sub-task × 3 sub-tasks × 20 tasks). Participants assigned to the non-curriculum condition followed the same task structure as the curriculum condition but with minor changes. In this condition, participants were told that the manufacturers only allowed them to play against the compositional slot machine. As a result, participants only interacted with one slot machine with rewards coming from the additive composition which results in a total of 100 trials (breakup: 5 trials per sub-task × 1 sub-task × 20 tasks)

Computational models

In this section, we describe the models that can perform the task with each model making a different assumption on how people could be approaching our task. A complete description of the models can be found in the SI.

Bayesian models

Under Gaussian assumptions, a Bayesian Mean-Tracker is often used to track a time-varying reward function $f_t(a)$ for option a on trial t. The mean is assumed to change over trials according to a Gaussian random walk:

$$f_{t+1}(a) \sim \mathcal{N}\left(f_t(a), \sigma_\zeta^2\right) \tag{6}$$

where $\sigma_\zeta = 0.001$. We also considered a variant of BMT model that can compose learned rewards called compositional BMT. This model follows the same setup as BMT but has its prior mean for the last sub-task initialized to the composition of learned means from the first two sub-tasks. We provide additional details about model learning and inference in the SI.

Gaussian Process Regression models learn a distribution over functions $f(a)$ defined by a mean function $\mu(a)$ and a covariance, or kernel function $k(a,a')$, where a and a' are arms of the bandit. The mean function defines the expected function value, while the covariance function controls the dependence between the function values for different inputs:

$$f(a) \sim \mathcal{GP}(\mu(a), k(a,a')) \tag{7}$$

$$\mu(a) = \mathbb{E}[f(a)] \tag{8}$$

$$k\left(a,a'\right) = \mathbb{E}\left[\left(f(a)-\mu(a)\right)\left(f\left(a'\right)-\mu\left(a'\right)\right)\right] \tag{9}$$

For the GP-based agent, we considered the GPR model with radial basis function (RBF) as the kernel. k_{RBF} allows the GPR model to generalize its learned value estimates depending on how (spatially) similar the options are to each other.

$$k_{\text{RBF}}\left(a,a'\right) = \exp\left(-\frac{1}{2}\left(a-a'\right)^\top \Theta^{-2}\left(a-a'\right)\right) \tag{10}$$

where Θ is the length scale hyperparameter. Like BMT, we assume that our GPR agent maintains a separate GP for each sub-task and by design, also cannot do any compositional inference.

As GPRs with appropriate priors can approximate the true generative model used for sampling the reward distributions in our task. We constructed a GPR model with compositionality built in, called compositional GPR. Such a model can compose the learned reward estimates from the first two sub-tasks and hence, reason compositionally on the third sub-task. We again assume the agent maintains a separate GP for each sub-task. We set the prior mean of the GPs corresponding to the first two sub-tasks to zero. The covariance function for the first sub-task is defined through a linear kernel k_{linear} defined as

$$k_{\text{linear}}\left(a,a'\right) = v a^\top a' \tag{11}$$

where v is the scale hyperparameter. While the covariance function of the second sub-task is defined through a periodic kernel k_{periodic} defined as

$$k_{\text{periodic}}\left(a,a'\right) = \exp\left(-\frac{2\sin^2\left(\pi \mid a-a' \mid /p\right)}{\eta^2}\right) \tag{12}$$

where p is a hyperparameter that determines the period length and η is a lengthscale hyperparameter.

The means and kernels for the final sub-task are obtained by composing the means and kernels from the first two sub-tasks[55]. For the first trial of the additive composition, the kernel is set to the mean of the learned linear and the periodic kernel from the two sub-tasks. The compositional additive kernel k_{additive} is defined as:

$$k_{\text{additive}}\left(a,a'\right) = 0.5(k_{\text{linear}}\left(a,a'\right) + k_{\text{periodic}}\left(a,a'\right)) \tag{13}$$

While the prior mean for additive composition is set to the mean of the previously learned mean functions from the linear and periodic sub-tasks. For the first trial of the change-point compositions, the kernel entries are set to that of the linear kernel if both arms belong to the linear function, the periodic kernel if both belong to the periodic function, and zero otherwise. The compositional change-point kernel $k_{\text{change-point}}$ is defined as:

$$k_{\text{change-point}}\left(a,a'\right)=k_{\text{linear}}\left(a,a'\right)\alpha_{\text{linear}}\left(a,a'\right)+k_{\text{periodic}}\left(a,a'\right)\alpha_{\text{periodic}}\left(a,a'\right) \tag{14}$$

where

$$\alpha_{\text{linear}}\left(a,a'\right)=\begin{cases}1 & \text{if } a,a'\in\{0,1,2\}\\ 0 & \text{otherwise}\end{cases} \tag{15}$$

α_{periodic} is defined analogously, giving a value of 1 whenever both arms' $a,a'\in\{3,4,5\}$. Note that we randomized whether the first three arms would belong to the linear or the periodic function. The prior mean in the change-point composition is set to the means learned in linear and periodic sub-tasks for the corresponding arms.

Meta-reinforcement learning

The version of RL^2 we use consists of a recurrent neural network (RNN) network followed by two linear networks that output a policy and a value estimate respectively[36,37]. We denote the joint vector of parameters of this model with $\mathbf{W}$. The network receives task-relevant observations o_t along with the action a_{t-1} and reward from the previous time step r_{t-1} as input and outputs a policy $\pi(a_t|\mathbf{h}_t,\mathbf{W})$ and a value estimate conditioned on the updated hidden state of the RNN. RL^2 is trained on samples from a task distribution $p(\omega)$ to find the policy that maximizes the sum of rewards in an episode of finite horizon H. The full objective function being optimized is shown in Equation 16:

$$\max_{\mathbf{W}}\mathbb{E}_{p(\omega)\prod p(r_t,o_{t+1}|a_t,\omega)\pi(a_t|\mathbf{h}_t,\mathbf{W})}\left[\sum_{t=1}^{H}r_t\right] \tag{16}$$

The particular resource constraint considered in RR-RL^2 is the description length of the meta-learned RL algorithm, which is defined as the number of bits required to store its parameters. Mathematically, this can be accomplished through a simple modification of Equation 16:

$$\begin{aligned}&\max_{\Lambda}\ \mathbb{E}_{q(\mathbf{W}|\Lambda)p(\omega)\prod p(r_t,o_{t+1}|a_t,\omega)\pi(a_t|h_t,\mathbf{W})}\left[\sum_{t=1}^{H}r_t\right]\\&\text{s.t. KL}[q(\mathbf{W}|\Lambda)\|p(\mathbf{W})]\leq C\end{aligned} \tag{17}$$

RR-RL^2 differs from RL^2 in two important ways. First, it uses a stochastic parameter encoding over neural network weights $q(\mathbf{W}|\Lambda)$ instead of a point estimate. Second, it places a constraint on the Kullback–Leibler (KL) divergence between $q(\mathbf{W}|\Lambda)$ and a prior $p(\mathbf{W})$, effectively limiting the number of bits that are needed to store the network's parameters and therefore the emerging reinforcement learning algorithm[56].

The network architecture of RL^2 and RR-RL^2 agents consisted of a gated-recurrent unit of size 128[57] followed by two linear layers that map hidden state-to-value function and policy respectively. The model implementations closely followed the implementation of Binz and Schulz[40]. We used a variational dropout prior[58] for RR-RL^2 and assumed that the encoding distribution factorizes into a set of independent normal distributions with learnable means and log-variances. Models were trained using a standard actor-critic loss at the end of each episode[59]. We used the ADAM optimizer[60] with a learning rate of 0.001 and trained for a total of 10^6 episodes with batches of size 32. RR-RL^2 relied on a dual gradient ascent procedure to enforce the constraint on the KL divergence[61]. We obtained gradients w.r.t. the parameters of the encoding distribution λ using the reparametrization trick[62]. We trained RR-RL^2 with description lengths between 10 and 10000 nats.

Modeling fitting and comparison

Bayesian models

The parameters of the Bayesian models were optimised endogenously for each participant, i.e., parameters of the model are chosen to maximise the likelihood of the data observed so far as in Schulz and colleagues[29]. We feed in the choice taken and reward received by participants from the previous trial and predict the expected reward and its uncertainty measure for all six options for the given trial. Note that the predictions are made after each trial conditioned on all data points up to that trial in the given task. The kernel parameters of these models are learned via gradient descent using the ADAM optimiser[60] for 100 iterations. The initial prior noise of these models was set to 0.001.

Following prior work[63,64], we use a variant of upper confidence bound sampling with an additional stickiness component as an action selection policy for both Bayesian models:

$$z(a_t|\beta,\tau,\lambda) = \beta\mu(a_t) + \tau\sigma(a_t) + \lambda\delta(a_t, a_{t-1}) \tag{18}$$

where $\delta(a_t, a_{t-1})$ takes the value of 1 if $a_t = a_{t-1}$ and 0 otherwise. This formulation includes uncertainty estimates for the learned values as an additional term to guide exploration. It has been shown to capture human behaviour well in function learning tasks[28,64] and also comes with performance guarantees[65].

The policy $p_{\text{Bayesian}}(a_t|\beta,\tau,\lambda)$ is then derived from these values using the softmax function:

$$p_{\text{Bayesian}}(i) = \frac{e^{z(a_t)}}{\sum_{j=1}^{K} e^{z(a_t)}} \quad for\ a_t = 0, 1, \ldots, 5 \tag{19}$$

The free parameters β, τ, and λ were fitted to human choices using a Bayesian model fitting procedure for each participant separately with the priors for parameters set to $\mathscr{N}(0,5)$. Model fitting was performed using the probabilistic programming toolbox PYMC3[66]. We used the marginal likelihood on the first trial of the compositional sub-task for model comparisons.

Meta-reinforcement learning

For modelling human choices, we assumed a mixture policy of the policy provided by the meta-reinforcement learning agent, a random policy, and a stickiness term:

$$p_{\text{RL}^2}(a_t \mid \varepsilon, \lambda) = (1 - \varepsilon - \lambda)\pi(a_t \mid \mathbf{h}_t) + \varepsilon|\mathscr{A}|^{-1} + \lambda\delta(a_t, a_{t-1}) \tag{20}$$

$$p_{\text{RR-RL}^2}(a_t \mid \varepsilon, \lambda, C) = (1 - \varepsilon - \lambda)\pi(a_t \mid \mathbf{h}_t, C) + \varepsilon|\mathscr{A}|^{-1} + \lambda\delta(a_t, a_{t-1}) \tag{21}$$

where C, ε and λ are free parameters, $|\mathscr{A}|$ denotes the number of available actions, and $\delta(a_t, a_{t-1})$ takes the value of 1 if $a_t = a_{t-1}$ and 0 otherwise. The marginal distribution $\pi(a_t \mid \mathbf{h}_t)$ was approximated with 10 samples from the encoding distribution.

We performed a grid search over the free parameters ε, λ and C and obtained a log-likelihood estimate for all pairs of parameters. ε and λ could take values between 0 and 1 with increments of 0.02, subject to the constraint that their sum is less than or equal to 1. The description length C could take values from 10 to $10,000$ in steps of 10. We assumed a uniform prior probability over these discretized parameter values, which allows us to compute the marginal log-likelihood for the first trial of the compositional sub-task as follows:

$$\log \sum_{\varepsilon}\sum_{\lambda}\sum_{C} \exp\left(\sum_{n=1}^{N} p_{\text{RR-RL}^2}(a_{11,n,i} \mid \varepsilon, \lambda, C)\right) - \log(N_C \cdot N_\varepsilon \cdot N_\lambda) \tag{22}$$

where N_C, N_ε, and N_λ correspond to the number of considered values for each parameter.

Model comparison

To obtain a quantitative measure of the goodness-of-fit to human choices, we conducted a Bayesian model comparison of all previously outlined models. We provide the full list of fitted parameters for each model in the SI. We measured the fit to human choices based on two metrics: posterior model frequency and exceedance probability[41]. The posterior model frequency measures how often a model offers the best explanation in the population, while the exceedance probability measures how likely it is that a given model is the most frequent explanation. We compute the metrics for model comparison using a Python implementation of the Variational Bayesian Analysis (VBA) toolbox [URL]. The toolbox requires us to provide log evidence – the marginal log-likelihood from the model fitting procedure in our case – for each model and participant, which we compute as previously described. For further details about this model comparison procedure see Rigoux and colleagues[41].

Acknowledgements

This work was supported by the Max Planck Society, the Volkswagen Foundation, the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.

Author contributions statement

A.J., M.B., and E.S. conceived the experiments, A.J. conducted the experiments, A.J. and T.S. analysed the results under M.B.'s, J.W.'s, and E.S.'s supervision. A.J. and M.B. wrote the main draft of the manuscript and J.W., T.S., and E.S. provided comments and reviewed the manuscript.

Additional information

The authors declare no competing interests.

References

1. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).
2. Carey, S. & Bartlett, E. Acquiring a single new word. *Reports on Child Lang. Dev.* (1978).
3. Schmidt, L. A. *Meaning and compositionality as statistical induction of categories and constraints*. Ph.D. thesis, Massachusetts Institute of Technology (2009).
4. Griffiths, T. L. Revealing ontological commitments by magic. *Cognition* **136**, 43–48 (2015).
5. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *arXiv preprint arXiv:1604.00289* (2016).
6. Frankland, S. M. & Greene, J. D. Concepts and compositionality: in search of the brain's language of thought. *Annu. review psychology* **71**, 273–303 (2020).
7. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
8. Franklin, N. T. & Frank, M. J. Generalizing to generalize: Humans flexibly switch between compositional and conjunctive structures during reinforcement learning. *PLoS computational biology* **16**, e1007720 (2020).
9. Biederman, I. Recognition-by-components: a theory of human image understanding. *Psychol. review* **94**, 115 (1987).
10. Collins, A. G. E. & Frank, M. J. Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition* **152**, 160–169 (2016).
11. James, W. *The Principles Of Psychology Volume II By William James (1890)* (Henry Holt and company, New York, 1890).
12. Kemp, C. Exploring the conceptual universe. *Psychol. Rev.* **119**, 685 (2012).
13. Xia, L. & Collins, A. G. Temporal and state abstractions for efficient learning, transfer, and composition in humans. *Psychol. review* **128**, 643 (2021).
14. von Humboldt, W. *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistigen Entwickelung des Menschengeschlechts* (Königliche Akademie der Wissenschaften, 1836).
15. Chomsky, N. *Aspects of the Theory of Syntax*, vol. 11 (MIT Press, 2014).
16. Griffiths, T. L. & Tenenbaum, J. B. Theory-based causal induction. *Psychol. review* **116**, 661 (2009).
17. Griffiths, T. L. Formalizing prior knowledge in causal induction. *The Oxf. Handb. Causal Reason.* 115 (2017).
18. Schulz, E. *Towards a unifying theory of generalization*. Ph.D. thesis, UCL (University College London) (2017).
19. Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M. & Gershman, S. J. Probing the compositionality of intuitive functions. Tech. Rep., Center for Brains, Minds and Machines (CBMM) (2016).
20. Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M. & Gershman, S. J. Compositional inductive biases in function learning. *Cogn. Psychol.* **99**, 44–79, DOI: 10.1016/j.cogpsych.2017.11.002 (2017).
21. Sanborn, A. & Griffiths, T. L. Markov chain monte carlo with people. In *Advances in Neural Information Processing Systems*, 1265–1272 (2008).
22. Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B. & Ghahramani, Z. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, 1166–1174 (2013).
23. Kumar, S., Dasgupta, I., Cohen, J., Daw, N. & Griffiths, T. Meta-learning of structured task distributions in humans and machines. In *International Conference on Learning Representations* (2020).
24. Amalric, M. *et al.* The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers. *PLoS computational biology* **13**, e1005273 (2017).
25. Piantadosi, S. T., Tenenbaum, J. B. & Goodman, N. D. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychol. review* **123**, 392 (2016).

26. Sablé-Meyer, M., Ellis, K., Tenenbaum, J. & Dehaene, S. A language of thought for the mental representation of geometric shapes. *Cogn. Psychol.* **139**, 101527 (2022).

27. Planton, S. *et al.* A theory of memory for binary sequences: Evidence for a mental compression algorithm in humans. *PLoS computational biology* **17**, e1008598 (2021).

28. Schulz, E., Konstantinidis, E. & Speekenbrink, M. Putting bandits into context: How function learning supports decision making. *J. experimental psychology: learning, memory, cognition* **44**, 927 (2018).

29. Schulz, E., Franklin, N. T. & Gershman, S. J. Finding structure in multi-armed bandits. *Cogn. Psychol.* **119**, 101261 (2020).

30. Saanum, T., Schulz, E. & Speekenbrink, M. Compositional generalization in multi-armed bandits. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43 (2021).

31. Lieder, F. & Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43** (2020).

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

33. Speekenbrink, M. & Konstantinidis, E. Uncertainty and exploration in a restless bandit problem. *Top. cognitive science* **7**, 351–367 (2015).

34. Rasmussen, C. E. Gaussian processes in machine learning. In *Summer school on machine learning*, 63–71 (Springer, 2003).

35. Schulz, E., Speekenbrink, M. & Krause, A. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *J. Math. Psychol.* **85**, 1–16 (2018).

36. Duan, Y. *et al.* Rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779* (2016).

37. Wang, J. X. *et al.* Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763* (2016).

38. Ortega, P. A. *et al.* Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030* (2019).

39. Binz, M. *et al.* Meta-learned models of cognition. *arXiv preprint arXiv:2304.06729* (2023).

40. Binz, M. & Schulz, E. Modeling human exploration through resource-rational reinforcement learning. *Adv. Neural Inf. Process. Syst.* **35**, 31755–31768 (2022).

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

42. Bhui, R., Lai, L. & Gershman, S. J. Resource-rational decision making. *Curr. Opin. Behav. Sci.* **41**, 15–21 (2021).

43. Correa, C. G., Ho, M. K., Callaway, F., Daw, N. D. & Griffiths, T. L. Humans decompose tasks by trading off utility and computational cost. *arXiv preprint arXiv:2211.03890* (2022).

44. Callaway, F. *et al.* Rational use of cognitive resources in human planning. *Nat. Hum. Behav.* **6**, 1112–1125 (2022).

45. Binz, M. & Schulz, E. Reconstructing the einstellung effect. *Comput. Brain & Behav.* 1–17 (2022).

46. Franklin, N. T. & Frank, M. J. Compositional clustering in task structure learning. *PLoS computational biology* **14**, e1006116 (2018).

47. Lake, B. & Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, 2873–2882 (PMLR, 2018).

48. Hupkes, D., Dankers, V., Mul, M. & Bruni, E. Compositionality decomposed: How do neural networks generalise? *J. Artif. Intell. Res.* **67**, 757–795 (2020).

49. Lake, B. M. Compositional generalization through meta sequence-to-sequence learning. *Adv. neural information processing systems* **32** (2019).

50. Kumar, S. *et al.* Using natural language and program abstractions to instill human inductive biases in machines. *Adv. Neural Inf. Process. Syst.* **35**, 167–180 (2022).

51. Dekker, R. B., Otto, F. & Summerfield, C. Determinants of human compositional generalization. *PsyArXiv preprint* (2022).

52. Peng, X. B., Chang, M., Zhang, G., Abbeel, P. & Levine, S. Mcp: Learning composable hierarchical control with multiplicative compositional policies. *Adv. Neural Inf. Process. Syst.* **32** (2019).

53. Andreas, J., Klein, D. & Levine, S. Learning with latent language. *arXiv preprint arXiv:1711.00482* (2017).

54. Dubey, R., Ho, M. K., Mehta, H. & Griffiths, T. Aha! moments correspond to meta-cognitive prediction errors. *PsyArXiv preprint* (2021).

55. Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J. & Zoubin, G. Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning* (2013).

56. Hinton, G. E. & 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*, 5–13 (1993).

57. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

58. Kingma, D. P., Salimans, T. & Welling, M. Variational dropout and the local reparameterization trick. *Adv. neural information processing systems* **28** (2015).

59. Mnih, V. *et al.* Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937 (2016).

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

61. Haarnoja, T. *et al.* Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905* (2018).

62. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

63. Borji, A. & Itti, L. Bayesian optimization explains human active search. *Adv. neural information processing systems* **26** (2013).

64. Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D. & Meder, B. Generalization guides human exploration in vast decision spaces. *Nat. human behaviour* **2**, 915–924 (2018).

65. Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. W. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE transactions on information theory* **58**, 3250–3265 (2012).

66. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in python using pymc3. *PeerJ Comput. Sci.* **2**, e55 (2016).