Compositional generalization in multi-armed bandits

Tankred Saanum[1, 2] **(tankred.saanum@gmail.com), Eric Schulz**[1]**, & Maarten Speekenbrink**[2]
[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2]Department of Experimental Psychology, University College London, London, UK

Abstract

To what extent do human reward learning and decision-making rely on the ability to represent and generate richly structured relationships between options? We provide evidence that structure learning and the principle of compositionality play crucial roles in human reinforcement learning. In a new multi-armed bandit paradigm, we found evidence that participants are able to learn representations of different reward structures and combine them to make correct generalizations about options in novel contexts. Moreover, we found substantial evidence that participants transferred knowledge of simpler reward structures to make compositional generalizations about rewards in complex contexts. This allowed participants to accumulate more rewards earlier, and to explore less whenever such knowledge transfer was possible. We also provide a computational model which is able to generalize and compose knowledge for complex reward structures. This model describes participant behaviour in the compositional generalization task better than various other models of decision-making and transfer learning.

Keywords: Compositionality; Reinforcement learning; Transfer learning; Gaussian Processes;

Introduction

Humans have a remarkable propensity for discovering structure in data. For instance, we can easily recognize that certain quantities, like global CO2 emissions, increase approximately linearly with time, or that the quality of certain fruits and vegetables varies periodically with the time of year. Moreover, having learnt such representations, we can combine and compose them to generate more sophisticated structures. For instance, if we know that a variable increases linearly over years, and periodically within years, we can combine this knowledge into a compositional representation of the statistical relationship between the variable and time.

Compositionality is argued to be indispensable to various parts of human cognition, such as language and reasoning (Hauser, Chomsky, & Fitch, 2002; Fodor, 1987), but its role in reinforcement learning (RL) has received less attention. Recent work shows that structure and function learning support reward prediction and guide exploration in RL tasks (Schulz, Franklin, & Gershman, 2020; Stojić, Schulz, P. Analytis, & Speekenbrink, 2020). The ability to combine such learnt representations compositionally could prove highly advantageous to performance in novel and complex RL situations.

Seeking to expound upon these ideas, we introduce the *compositionally-structured bandit* task, a paradigm for studying compositional generalization and transfer learning in a class of RL tasks in which an agent needs to sequentially choose between options (the "arms" of the bandit), balancing exploration and exploitation in order to accumulate as much reward as possible. Crucially, in our paradigm, certain reward functions the agent encounters are *compositions* of reward functions encountered previously, allowing the agent to gain more rewards by harnessing the appropriate compositional inductive biases.

To foreshadow our results, we found substantial evidence that participants composed knowledge of previously learnt reward structures to make generalizations about rewards in novel situations. This compositional knowledge transfer afforded participants the ability to make more informed decisions in these novel situations, allowing them to focus their decisions and exploration on more rewarding options and hence accumulate more rewards earlier on. We also propose a novel computational model that combines symbolic reasoning, embodied in a generative grammar, with statistical inference, embodied in *Gaussian process* regression, which can reproduce human behaviour in our task. Ultimately, our results suggest that the principles of compositionality may play a crucial role in human reward learning and decision-making.

The compositionally-structured bandit

In traditional bandit tasks, the agent selects at each trial t an option/arm $a_t \in \mathcal{A}$ from the choice set $\mathcal{A}$, which produces a reward r_t drawn from an option-specific reward distribution. In structured bandit tasks (Schulz et al., 2020; Stojić et al., 2020), each arm a is described by a set of features $\mathbf{x}_a$, and rewards are drawn from a joint distribution that is governed by a latent function f:

$$r_t = f(\mathbf{x}_{a_t}) + \varepsilon_t \tag{1}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. As such, rewards are determined by a latent function defined over the features of the arms in the choice set. Critically, this function defines a structural relation between the features and rewards. For instance, rewards may increase linearly with certain features of the arms, such as their spatial position, their color, and so forth. Learning the structures that govern the rewards can significantly improve performance, as it allows the agent to generalize from past observations to novel, unexplored options of the choice set.

The compositionally-structured bandit extends this paradigm by allowing rewards to be governed by multiple latent reward functions, where each reward function is associated with a set of contextual features $\mathbf{f}$. In such multi-task or contextual bandit problems, the expected rewards associated with each arm changes depending on the contextual features $\mathbf{f}$, i.e. which latent function currently governs its reward distribution. As such, the agent is tasked with learning a separate reward function for each context. Finally, in our paradigm the features of a context may be *composed* of other features which the agent explored and learned about in the past. That is, certain contexts will contain the features of multiple other contexts which the agent encountered in the past. We call such contexts compositional. Crucially, in our setup the latent reward function in these contexts is always an additive composition of the latent functions governing rewards in the corresponding constituent contexts. This way, the agent can make informed predictions about the reward structure in compositional contexts by composing the reward functions learnt in the relevant prior contexts.

We implemented this bandit task in a game where participants had to sequentially choose which dish (arm) to offer to alien customers in order to maximise the customers' payment (reward). Each dish was made up of a specific *amount* of two visually distinct ingredients (features), $\mathbf{x}$. Alien customers were adorned with visual attributes, which reflected the contextual features $\mathbf{f}$. Rewards were defined by context-specific reward functions over the two-dimensional features: $r_t = f(\mathbf{x}_t, \mathbf{f}_t) + \varepsilon_t$. Crucially, the reward functions for some contexts were compositions of the reward functions of other contexts. The game was structured in four rounds, each containing several contextual reward functions (both compositional and not), for which participants had a set amount of trials to select arms. With this task, we set out to investigate how humans explored and exploited options in a task which allowed both for structure learning and compositional generalization. We hypothesized that participants would compose representations of reward-structures learnt in the past to make informed decisions when the latent function and contextual features were compositional. As such, we expected the rewards obtained on the first trial of compositional contexts to be higher than those of non-compositional contexts. Moreover, we reasoned that if participants composed reward functions from past contexts, then uncertainty around the compositional context would be significantly reduced, and the need for exploration would decrease. We therefore hypothesized that participants would explore less in compositional contexts.

Method

Participants

We recruited 47 participants (22 female, $M_{age} = 28.3$, $SD_{age} = 7.5$) through Prolific. All participants had an approval rate of 95% or more, were fluent English speakers and had no color vision deficiencies. To improve the quality of the data, participants had to complete a tutorial before beginning the task. Participants were rewarded a base payment of £1.92 and a performance-dependent bonus payment of £1 on average. It took participants 22.9 minutes to complete the task on average ($SD_{time} = 6.9$).

Task and procedure

Participants gave their informed consent and were instructed that they would play a game where their job was to combine two types of ingredients to make and sell food items to aliens. The aliens would then give money for the served food and the amount of money they gave depended on how much they liked the food they were served. Participants were also informed that the aliens had different colored symbols on their bellies that signal their food preferences, and that aliens with similar features had similar preferences.

On each trial, participants were presented with an alien customer (contextual cue), and then selected a dish to serve from a two-dimensional ingredient space, in which each dimension corresponded to the amount (between 0 to 10) of an ingredient. Formally, each possible dish from the 11×11 grid representing the feature space was a reward-generating arm in a contextual bandit task. Participants were also informed how many trials were left in the current round.

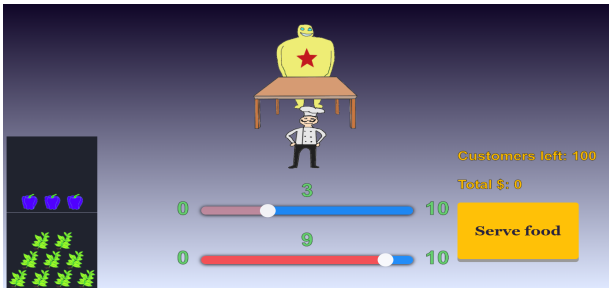

Figure 1: Screenshot from the task. The option features correspond to amounts of ingredients, which can be selected by adjusting the sliders.

For each context, the reward function was either a non-compositional linear or periodic function defined over a single dimension in the input space, or an additive composition of such functions (see Figure 2). Which latent function currently governed the reward distributions was determined by the contextual cues (i.e. the colored symbols on the alien's belly). The non-compositional functions were accompanied with either a star or triangle symbol, rendered in either red or blue. The symbol type always matched the type of reward function (linear or periodic), whereas the symbol color always matched which input dimension this function was defined over (i.e. the first or second ingredient). Allocation of functions and dimensions to symbol types and colors was randomized for each participant. For contexts featuring two symbols, the latent reward function was a composition of the functions related to each symbol in isolation. In total, participants were tested on 10 unique reward functions, six of which were compositional (see Table 1).

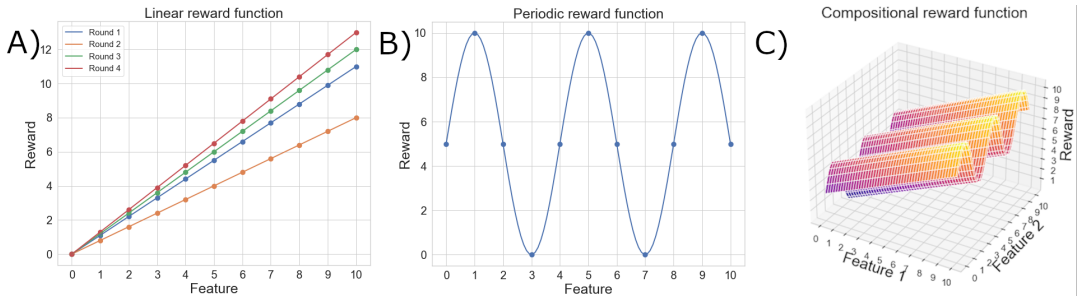

Figure 2: **A**: The linear reward functions used in the four rounds. **B**: The periodic reward function, reused for all rounds. **C**: The compositional, linear-periodic reward function from Round 1.

Round	Latent function	Trials
1	$Lin(X)$, $Per(Y)$, $Lin(X)+Per(Y)$	40, 40, 20
2	$Lin(Y)$, $Per(X)$, $Lin(Y)+Per(X)$	20, 20, 20
3	$Lin(X)$, $Per(Y)$, $Lin(Y)$, $Per(X)$, $Lin(X)+Per(X)$, $Lin(Y)+Per(Y)$	10, 10, 10, 10, 10, 10
4	$Lin(X)$, $Per(Y)$, $Lin(X)+Lin(Y)$, $Per(X)+Per(Y)$, $Lin(X)+Per(X)$, $Lin(Y)+Per(Y)$, $Lin(X)+Per(Y)$, $Lin(Y)+Per(X)$	10, 10, 10, 10, 10, 10, 10, 10

Table 1: The four rounds, and the latent reward functions they feature. *Lin* denotes the linear function, and *Per* the periodic function. The X's and Y's point to the dimension over which the function was defined. The trial column indicates how many trials participants had to select arms for the respective functions in that round. Reward functions appeared in the order they are shown in the table, except for in round 4, where the order was randomized. When a function only mentions one dimension (e.g. $Lin(x)$), the other dimension (Y) is unrelated to reward.

Behavioral Results

In all rounds, participants gained significantly more rewards than the chance levels of those rounds (round 1, $t(146.56) = 36.38$, $p < .001$, round 2 $t(146.05) = 27.29$, $p < .001$, round 3 $t(312.14) = 32.58$, $p < .001$, round 4 $t(392.91) = 35.32$, $p < .001$), indicating that they were able to learn about and exploit the latent reward functions.

Transfer learning

To test for transfer learning, we analysed the reward obtained on the first trial of each context. Compositional generalization is indicated by a higher reward for the first encounter with a compositional context than for the first encounter with the constituting simple contexts. As numerical rewards varied with reward functions, we first normalized all obtained rewards as the fraction of the maximum attainable reward for that context. We then used a linear mixed-effects model predicting these normalized rewards from reward function type and round (expressed as orthogonal contrast codes). The model also included participant-specific random intercepts and slopes to account for individual differences in learning.

Rewards obtained on the first compositional trials were significantly higher than rewards obtained on the first trials of the simple functions $t(846) = 6.35$, $p < .001$. On average, participants scored 17.15% higher on the first trial of compositional contexts than on the first trial of simple contexts. Before having been given the chance to learn from it directly, participants were significantly more likely to select an optimal arm on the *first* trial in the compositional contexts than chance level (computed as the number of optimal arms divided by the total number of arms), both in the first ($t(46) = 4.05$, $p < .001$) and the second round ($t(46) = 5.24$, $p < .001$). Interestingly, on the first trial of the compositional context in round 1, participants had not yet had the opportunity to learn that compositional contexts had compositional reward functions, indicating that their employment of compositional generalization reflected an a priori inductive bias, rather than something they learnt through trial and error.

We also investigated whether participants *improved* in harnessing compositionality to make informed decisions in compositional contexts. To assess whether there was such a learning-to-learn effect (Harlow, 1949) for compositional inference, we tested whether mean rewards for the first compositional trial in round 1 were significantly different from those in round 2. Indeed, a t-test revealed that mean rewards for the first compositional trial in round 2 were significantly higher than those of round 1, $t(46) = 2.51, p = .02$. Furthermore, more participants selected the optimal arm on the first trial of the compositional context in round 2 than in round 1 (40.4% percent compared to 29.8% percent), though this difference was not statistically significant, $t(46) = 1.3, p = .2$. Though the number of participants who selected an optimal arm on the first trial of the compositional context in round 3 was not significantly different than chance $t(93) = 1.06, p = .2$, in round 4 significantly more chose an optimal arm than chance on the first trial of contexts containing functions composed of two linear functions $t(46) = 4.76, p < .001$, or two periodic functions $t(46) = 4.82, p < .001$.

Exploration

To assess how participants explored in the task's different contexts, we sought to predict exploration with a mixed effects model, using the same predictors as the transfer learning model, and random intercepts and slopes for the effect of round. We operationalized relative exploration through the Shannon entropy of the distribution over participants' choices over the two-dimensional feature space in each context. Informally, the entropy of a distribution quantifies its unpredictability. As a distribution approaches uniform, its entropy

increases, and vice versa. Consequently, in contexts where participants explore a larger portion of the choice set, the corresponding choice distribution will be more uniform, and entropy will be high. As such, entropy will be low when participants exploit more, or employ more strongly guided exploration.

Participants explored less in compositional contexts compared to simple contexts: There was a significant difference between the entropy of participants' choice distributions in the compositional and simple contexts, $t(751.03) = -8.71$, $p < .001$, indicating that participants explored more for simple functions, and exploited more for compositional functions.

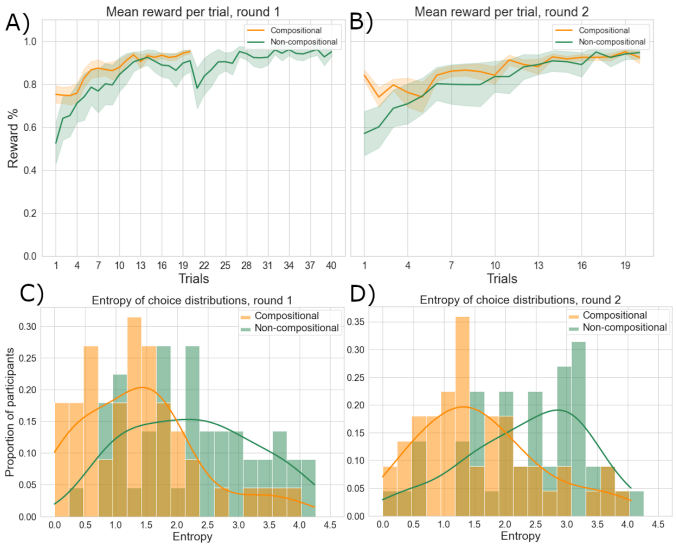

Figure 3: **A, B**: Mean rewards per trial obtained for compositional and average of non-compositional reward functions in round 1 and 2. **C, D**: Entropy histograms for compositional and average of non-compositional reward functions in round 1 and 2. Plotted lines represent the kernel density estimate of the histograms.

Model-based analysis

The behavioral results indicate participants were able to transfer their knowledge between contexts, composing new reward functions by combining simpler reward functions in a productive fashion. To account for this, we now propose a computational-level model which is able to learn and compose such structures through Bayesian inference and knowledge transfer over a grammar of functions.

Earlier work has modelled human function learning as Gaussian process (GP) regression (Griffiths, Lucas, Williams, & Kalish, 2009; Schulz et al., 2020). This is a non-parametric Bayesian method for inferring functions from data, and when coupled with a decision strategy such as Upper Confidence Bound (UCB) sampling, also provides a good account of human behaviour in multi-armed bandit tasks (Stojić et al., 2020; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018).

Gaussian Processes

A Gaussian process defines a distribution over functions, such that for any set of input points $\mathbf{x}_1, ..., \mathbf{x}_n$, the outputs $f(\mathbf{x}_1), ..., f(\mathbf{x}_n)$ follow a joint (multivariate) Gaussian distribution, $f \sim \mathcal{GP}$. In our case, the input points $\mathbf{x}$ are the features of the arms (ingredient combinations) in the contextual bandit task, and the outputs $f(\mathbf{x})$ are the rewards.

A GP is defined by a mean function $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and a covariance function $k(\mathbf{x}, \mathbf{x}')) = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$. The latter is also known as the kernel and defines how the random outputs of any two input points covary. The posterior distribution over the outputs, given observations $\mathcal{D}_n = \{\mathbf{X}_n = [\mathbf{x}_1, ..., \mathbf{x}_n], \mathbf{y}_n = [y_1, ...y_n]\}$, is also a GP with mean and kernel function

$$m_{\text{post}}(\mathbf{x}) = \mathbf{k}^T (\mathbf{K} + \sigma^2)^{-1} \mathbf{y}^T \quad (2)$$

$$k_{\text{post}}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}^T \ (\mathbf{K} + \sigma^2)^{-1} \mathbf{k} \quad (3)$$

(Schulz, Speekenbrink, & Krause, 2018) where $k(\mathbf{x}, \mathbf{x}')$ is the kernel function, $\mathbf{k}$ is the kernel matrix containing the prior covariance between testing and observed input points, and $\mathbf{K}$ is the kernel matrix containing the covariance between all observed input points. Consequently, in our task, the rewards generated at each arm are modelled as normally distributed random variables in a GP. To make predictions about rewards in a given context c, we derive the posterior GP using the rewards already observed in c, and use its mean function $m(\mathbf{x})$ to make predictions.

The kernel is central to GP regression: There are several kinds of kernels, each specifying different structures which are imposed on the functions modelled by the GP, and each encoding assumptions about the functions' structure, such as linearity and periodicity, smoothness and noise. As such, how a GP interpolates and extrapolates from observations is determined by its kernel. Another crucial property of these positive-definite GP kernels is that they are closed under addition and multiplication, such that if k is a kernel function and k' is a kernel function, then so is $k + k'$ and $k \times k'$ (Duvenaud, 2014). Consequently, there is an infinite set of kernels available to model the covariance structure of a function. This makes GPs able to express a vast range of rich functional forms. Unfortunately, this also complicates the task of selecting an appropriate kernel for a particular regression problem, as there is an infinite set of arbitrarily complex candidate kernels to select from, each of which will have a different likelihood of generating the data.

Exploiting the compositional properties of kernels, we rely on a compositional kernel grammar (Duvenaud, 2014; Janz, Paige, Rainforth, van de Meent, & Wood, 2016) to solve this problem for the various reward structures participants encounter. The kernel grammar is a generative model of covariance functions which probabilistically produces kernels. In our approach, it starts by sampling a kernel from a base set containing the standard kernels used in the literature, namely the linear, periodic, and radial basis kernels (Duvenaud, 2014), $\mathcal{B} = \{k_{Lin}, k_{Per}, k_{RB}\}$, and recursively expands this kernel through a sequence of steps in which it samples a new kernel $k \sim \mathcal{B}$ and either adds or multiplies it with the current kernel. At each step, there is a probability γ that

the grammar stops and returns the current production. As such, the parameter γ controls the productivity of the grammar, and the complexity of the kernels being produced.

As this grammar implicitly defines a prior over kernel functions, we seek to approximate the posterior over kernels, embodying the hypothesized structure of the reward function being modelled, given the observations. We do this by first sampling 100 kernels from the grammar and computing the corresponding posterior GPs. We then obtain each GP's marginal likelihood, using their respective kernel in our hypothesis set, and compute the posterior probability of these kernels, given the reward data observed.

$$p(k \mid \mathcal{D}) \propto p(\mathcal{D} \mid k)p(k) \tag{4}$$

With this posterior distribution we compute a final posterior GP, which is the sum of all posterior GPs (using their respective kernels), weighted by their posterior probability. We rely on this procedure to capture the structure of the reward functions we tested participants on in our experiment.

Transfer learning

For the model to be able to compose reward structures from past contexts and tasks, we equipped it with what has been referred to as a Neural Dictionary (Pritzel et al., 2017). This dictionary consists of a set of keys, which are vectors $\mathbf{f}_c$ encoding the features of encountered contexts c, and corresponding values which are the posterior GPs learnt for c. Upon computing a posterior GP for a context, as per the last section, it writes an entry into the dictionary whose key is the feature vector of the context, and whose value is this posterior GP. If the context has been visited previously, it overwrites the old posterior GP with the new one. Crucially, whenever the model encounters a context whose reward function is composed of two previously seen functions, we ask it to transfer knowledge from previously explored contexts. This is achieved by computing the similarity between the current context c and all previously seen contexts c', $\kappa(c,c')$ where $\kappa(\cdot,\cdot)$ is a similarity measure. This assigns a similarity score $0 \leq \kappa(c,c') \leq 1$ to each c' in the dictionary, which we normalize by the total similarity. We found cosine similarity to be a suitable measure for our task

$$\kappa(c,c') := \frac{\mathbf{f}_c \cdot \mathbf{f}_{c'}}{\|\mathbf{f}_c\| \cdot \|\mathbf{f}_{c'}\|}. \tag{5}$$

With these similarity scores, we derive a new GP which is the sum of the posterior GPs stored in the dictionary, weighted by their similarity to the current context c_*.

$$\mathcal{GP}_* = \sum_i^N \mathcal{GP}_i \left(\frac{\kappa(c_*, c_i)}{\sum_j^m \kappa(c_*, c_j)} \right) \tag{6}$$

This not only allows the model to use informed priors about the current context's reward function based on its similarity to tasks encountered in the past, but also to *compose* previously learnt representations of the reward structure by adding them together, if the contexts in which these representations were learnt are similar to the current context.

Choice probabilities

The model derives a GP with a mean vector $m(\mathbf{x})$, describing the predicted rewards at each arm, and a covariance function $k(\mathbf{x},\mathbf{x}')$, with $\sigma(\mathbf{x}) = \sqrt{k(\mathbf{x},\mathbf{x})}$ reflecting the uncertainty of its predictions. We use both of these components to devise a decision strategy for the model. In particular, we evaluate the quality of each arm $Q(\mathbf{x})$ using the Upper confidence bound sampling (UCB) algorithm (Sutton & Barto, 2018)

$$Q(\mathbf{x}) = m(\mathbf{x}) + \beta\sigma(\mathbf{x}), \tag{7}$$

where β is a parameter controlling how reducing uncertainty should be traded off against exploiting higher-rewarding arms. As such, this strategy attempts to strike a balance between pure exploration and pure exploitation strategies, and is a solution to the exploration-exploitation trade-off. We convert the arms' Q-values to choice probabilities using a softmax function (discarding the temperature parameter τ)

$$P(\mathbf{x}) = \frac{\exp(Q(\mathbf{x}))}{\sum_i^N \exp(Q(\mathbf{x}_i))} \tag{8}$$

Results

We estimated how likely our model, which we will refer to as the *GP-grammar* model, was to produce participants' choices on the first trials of the compositional contexts. The complexity penalty γ for the kernel grammar was set to 0.8, and the exploration parameter β for the UCB strategy was set to 1.96 (reflecting the 95% confidence interval of the estimated reward). We compared the GP-grammar model to several alternatives: a random model which assigns to all arms a uniform probability of being selected $P(\mathbf{x}) = 1/121$, three lesioned versions of the GP-grammar model, employing only a linear, periodic, or RBF kernel, respectively, but retaining the Neural dictionary for transfer learning, and lastly a Universal Value Function Approximator (UVFA) (Schaul, Horgan, Gregor, & Silver, 2015). The UVFA is a state-of-the-art transfer learning model which, in our task, learns rewards both across options and contexts. In our approach, the UVFA took the form of a GP with an RBF kernel, using both option features and context features to learn and predict rewards. Since the RBF kernel is universal (Schölkopf & Smola, 2002), the UVFA and the lesioned RBF models could recover the compositional ground truth. The advantage of the GP-grammar model, however, lies in its ability to elicit the appropriate priors about the latent structure more strongly and with less data by performing Bayesian inference. Crucially, the models were tested on choices made before participants had observed any reward function values for this context. As such, any computational model that does not transfer knowledge from prior contexts on these trials will make identical predictions to the random model on these trials as well.

To estimate the GP-grammar model's performance in reproducing participants' choices, we sequentially derived the posterior GP for each context encountered by participants as described in the last section, conditioning the corresponding

GP on all input-output points that participant had observed for that context. We endowed each context with a one-hot encoded vector, encoding whether the context featured a star, a triangle, or both, and whether the symbol(s) were blue or red, respectively. In contexts whose latent reward function was compositional, we computed the contextually informed GP, based on the cosine similarity between the relevant context vectors, as per equations (5) and (6). Making use of the mean and covariance function of this GP, we computed the Q-values for all possible arms and converted them to choice probabilities, using (7) and (8) respectively. The lesioned models were trained the same way, and the UVFA was simply conditioned on all previously seen observations, including the relevant context vectors as features. For each participant, we obtained the average probability of generating their choices for all models, and summed up the models' log likelihoods across participants. With these quantities we computed the posterior probability of the models, assuming a uniform prior, as well as McFadden's pseudo-R^2 values (McFadden et al., 1973), $R^2 = 1 - \frac{\mathcal{L}(M)}{\mathcal{L}(M_{random})}$, quantifying the degree to which a model explains the variance over and above chance, where $R^2 = 0$ correspond to chance levels, and $R^2 = 1$ corresponds to a model infinitely more accurate than chance.

We found the GP-grammar model was substantially more likely to reproduce participants' choices than the other models, obtaining a posterior probability of $P(M \mid y) > 0.999$, and an R^2 score of 0.29, also higher than all alternative models (UVFA: $R^2 = 0.04$, linear kernel: $R^2 = 0.23$, periodic kernel: $R^2 = 0.05$, RBF kernel: $R^2 = 0.05$). That the GP-grammar model outperformed both the lesioned and UVFA models suggests that human compositional generalization in contextual bandit tasks not only relies on the ability to discover sophisticated reward structures, but also on the the ability to compose such structures.

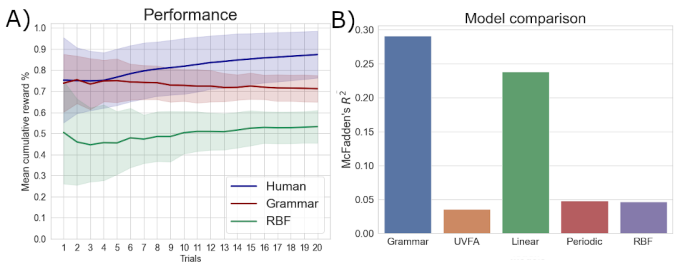

Figure 4: **A**: Participant performance in the compositional context of the first round compared to the performance of our full model ("Grammar") and that of a GP using an RBF kernel. **B**: McFadden's R^2 for all models. Our full model is denoted by "Grammar".

Moreover, our results suggest that this propensity relies both on the ability to discover richly structured representations of how rewards are distributed in a choice space and on being able to compose and combine such representations when contextual features call for it. As such, our Bayesian grammar-based approach for discovering viable covariance functions for GPs combined with a compositional transfer learning mechanism, presents itself as a suitable, Bayesian model of generalization in compositional bandit tasks.

Discussion

We assessed the extent to which humans harness compositionality to support reward learning and decision-making in RL settings. Our results indicate that participants were able to learn representations of the abstract structures governing how rewards were generated, and more importantly, make informed, compositional generalizations from these simpler representations. This is indicated by first rewards being significantly higher for contexts containing compositional latent functions: Since the compositional functions were always preceded by the simple functions from which they were composed, participants could make informed predictions about rewards in compositional contexts before having observed any input-output pairs. That participants explored less in compositional contexts offers further compelling evidence for this transfer of knowledge: By harnessing compositionality in novel contexts, participants could sidestep the need for exploration, and compose representations of past reward functions to select higher-rewarding options earlier. Ultimately, these results suggest that compositionality, a principle commonly invoked in linguistics and cognitive science to account for the productivity and systematicity of human cognition (Fodor, 1987; Lake, Salakhutdinov, & Tenenbaum, 2015), should be central in theories of human RL and decision-making as well.

We also developed a novel computational model able to reproduce core aspects of the compositional generalization observed in the behavioural data. This model conceptualizes reward-structure learning as GP regression, where the kernel embodying the latent structure is discovered through Bayesian inference over a set of compositional kernels produced by a generative grammar. The ability to compose such representations is conceptualized as similarity-based knowledge transfer, in which a novel representation is constructed as an additive composition of prior learnt representations, weighted by the similarity between the prior and present contexts. This model was substantially more likely to generate participant choices on the first trial of the compositional contexts than lesioned counterparts and another transfer learning model. In the end, our modelling results suggest that structure learning in humans may be supported by symbolic, grammar-like computations, and that contextual similarity judgements underpin how humans compose latent structures.

One current caveat is that we only tested participants' propensity for compositional knowledge transfer in settings where the latent function was an *additive* composition of two simpler functions. To gain further evidence for our hypotheses, in future work, we aim to test whether these behavioural effects persist in tasks with more complex compositional structures, such as multiplicative or change-point functions (Duvenaud, 2014), and extend our model to be able to generalize about such compositions as well.

References

Duvenaud, D. (2014). *Automatic model construction with Gaussian processes* Unpublished doctoral dissertation. University of Cambridge.

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.

Griffiths, T. L., Lucas, C., Williams, J., & Kalish, M. L. (2009). Modeling human function learning with Gaussian processes. In *Advances in neural information processing systems* (pp. 553–560).

Harlow, H. F. (1949). The formation of learning sets. *Psychological review*, *56*(1), 51.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Janz, D., Paige, B., Rainforth, T., van de Meent, J.-W., & Wood, F. (2016). Probabilistic structure discovery in time series data. *arXiv preprint arXiv:1611.06863*.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior.

Pritzel, A., Uria, B., Srinivasan, S., Puigdomènech, A., Vinyals, O., Hassabis, D., … Blundell, C. (2017, March). Neural Episodic Control. *arXiv:1703.01988*.

Schaul, T., Horgan, D., Gregor, K., & Silver, D. (2015, 07–09 Jul). Universal value function approximators. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (Vol. 37, pp. 1312–1320). Lille, France: PMLR.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA, USA: MIT Press.

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

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16.

Stojić, H., Schulz, E., P. Analytis, P., & Speekenbrink, M. (2020, March). It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *Journal of Experimental Psychology: General*, *149*(10), 1878–1907.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018, December). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924.