Probing Compositional Inference in Natural and Artificial Agents

Akshay K. Jagadish
Max Planck Institute for Biological Cybernetics
Tübingen, Germany
akshay.jagadish@tue.mpg.de

Tankred Saanum
Max Planck Institute for Biological Cybernetics
Tübingen, Germany

Jane X. Wang
DeepMind
London, United Kingdom

Marcel Binz
Max Planck Institute for Biological Cybernetics
Tübingen, Germany

Eric Schulz
Max Planck Institute for Biological Cybernetics
Tübingen, Germany

Abstract

People can easily evoke previously encountered concepts, compose them, and apply the result to novel contexts in a zero-shot manner. What computational mechanisms underpin this ability? To study this question, we propose an extension to the structured multi-armed bandit paradigm, which has been used to probe human function learning in previous works. This new paradigm involves a learning curriculum where agents first perform two sub-tasks in which rewards were sampled from differently structured reward functions, followed by a third sub-task in which rewards were set to a composition of the previously encountered reward functions. This setup allows us to investigate how people reason compositionally over learned functions, while still being simple enough to be tractable. Human behavior in such tasks has been predominantly modeled by computational models with hard-coded structures such as Bayesian grammars. We indeed find that such a model performs well on our task. However, they do not explain how people learn to compose reward functions via trial and error but have, instead, been hand-designed to generalize compositionally by expert researchers. How could the ability to compose ever emerge through trial and error? We propose a model based on the principle of meta-learning to tackle this challenge and find that – upon training on the previously described curriculum – meta-learned agents exhibit characteristics comparable to those of a Bayesian agent with compositional priors. Model simulations suggest that both models can compose earlier learned functions to generalize in a zero-shot manner. We complemented these model simulations results with a behavioral study, in which we investigated how human participants approach our task. We find that they are indeed able to perform zero-shot compositional reasoning as predicted by our models. Taken together, our study paves a way for studying compositional reinforcement learning in humans, symbolic, and sub-symbolic agents.

1 Introduction

Every day we bring together concepts from seemingly different settings to bear on problems where they have not been used before. To illustrate, let us take the example of reading trends in COVID vaccination numbers. Figure 1 shows the vaccination numbers in Germany from December 2020 to March 2021. You can clearly identify a pattern in the trajectory of these numbers. They increase roughly exponentially (indicated by the black dots) but with a periodic rise and fall every week. If you were shown the trend until the first week of March and asked to extrapolate to the second week (in the red box), you would likely predict something close to the true trend. How do we accomplish this?

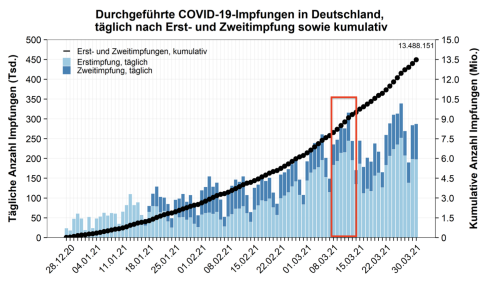


Figure 1: COVID vaccination numbers

The aforementioned example involves two aspects of learning that have been studied extensively in cognitive science but mostly in isolation [1, 2, 3]. First, function learning (i.e., learning the exponential and periodic function), and second, compositional reasoning over learned function (i.e., combining the exponential and periodic function to extrapolate the trend for next week). Recent work of [2], for instance, has investigated the first aspect. They introduced a structured multi-armed bandit (sMAB) task, where rewards for different options followed a latent correlation structure, to study how people explore in structured environments. The latent reward functions were either linear or periodic (structured condition) or set to random uncorrelated values (unstructured condition). Participants were told to earn as many reward points as possible but were not provided any information regarding the underlying reward functions. They found that participants could pick up on the latent functions in the structured condition after a few rounds and use it to guide their exploration in later rounds.

While the study of [2] provides insights into how people learn latent reward functions, it does not tell us how they re-use earlier learned functions to compose new functions that were never encountered before. We attempt to close this gap in the present article. To this end, we extended the paradigm of [2] by introducing a learning curriculum. Participants in our task first performed two sub-tasks in which rewards were sampled from differently structured reward functions, followed by a third sub-task in which the rewards were set to a composition of the previously sampled functions. We were primarily interested in the following questions:

1. Do people compose previously learned functions in the final sub-task in a zero-shot manner?
2. Which computational models capture people's compositional reasoning abilities?

We demonstrate that people can indeed learn to compose new functions in a zero-shot manner. Historically, human behavior in such tasks has been modeled using algorithms with hard-coded structures, such as Bayesian grammars [4]. We find that these models also perform well in our task. However, they cannot answer the questions of where these structures come from and how they are acquired in the first place. To address these questions, we propose an alternative model based on the principle of meta-learning [5]. When trained on an appropriate curriculum, this model exhibits characteristics comparable to those of a Bayesian agent with compositional priors. We briefly discuss the advantages and disadvantages of the two modeling approaches and suggest that they can be used to understand distinct aspects of human cognition.

2 Methods & Results

2.1 Compositional Multi-Armed Bandit Task

To probe compositional reasoning in humans and artificial agents, we developed a task based on earlier work of [2]. Each task consists of three sMAB sub-tasks, where rewards follow a function that is dependent on the spatial position of arms. In each sMAB, rewards were determined as follows:

$$r_t = f(a_t) + \epsilon_t \tag{1}$$

where t denotes the time-step, $a_t \in \{0, \ldots, 5\}$ the action taken in time-step t, and $\epsilon_t \sim \mathcal{N}(0, 0.3)$ is an additive noise term.

The first two sub-tasks involved latent functions from either the linear or the periodic family:

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

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

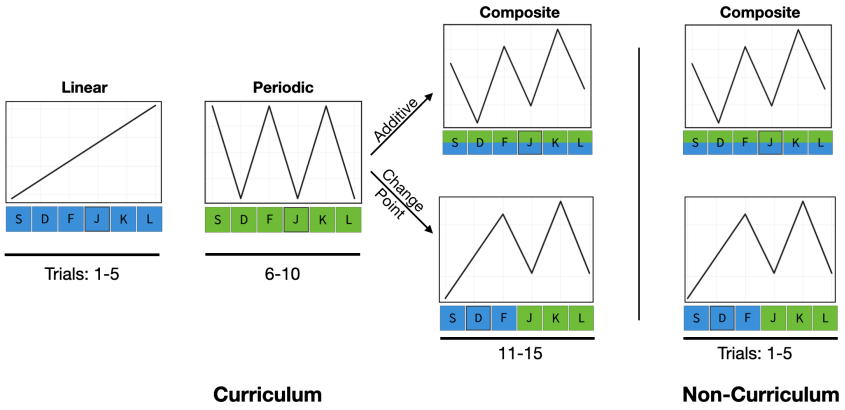


Figure 2: **Task Overview**. Example tasks both compositional rules and conditions.

Latent functions in the final sub-task were composed from the functions encountered in the two earlier sub-tasks. We considered both additive and change-point compositions:

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

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

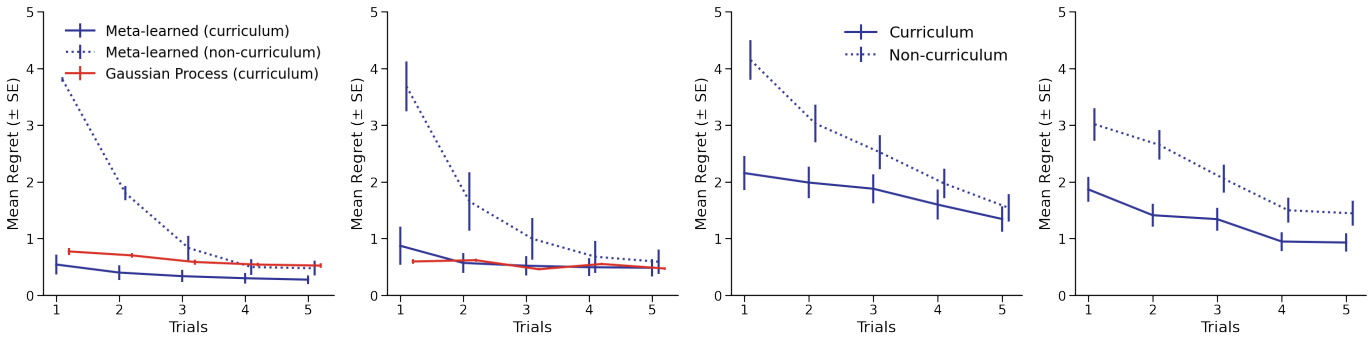
The order of composition was randomized in the change-point function. The length of each sub-task was set to 5 trials, leading to 15 overall trials per task. Note that the number of trials per sub-task is less than the number of available options. This prevents an agent from exhaustively trying out all options and forces it to generalize based on the underlying function. Figure 2 shows an example for both the additive and change-point condition. To have a comparison, we also consider a condition without curriculum. In this non-curriculum condition, the agent did not interact with the first two sub-tasks and instead directly observed the composite function from the final sub-task (as illustrated on the right in Figure 2).

2.2 Computational Models

The goal of an agent used to model our task is to maximize the total amount of obtained rewards. We outline two models for achieving this goal. The first is a Bayesian model that incorporates knowledge of priors via the Gaussian Process (GP) regression model; the second is a recurrent neural network model that is meta-learned based on repeated interactions with randomly sampled tasks.

Gaussian Process Model: We implemented a GP regression model as the Bayesian model for our task as they have been successful at capturing human function learning in previous works [4, 2]. The latent reward functions were learned using a linear kernel for the first sub-task, a periodic kernel for the second sub-task, and a kernel using a composition of the linear and periodic kernel for the last sub-task. For additive compositions, the compositional kernel was the sum of a linear and periodic kernel, and the prior mean was set to the mean of the previously learned functions from the linear and periodic sub-tasks. For change-point compositions, the kernel entries were set to that of the linear kernel if both arms belonged to the linear function, the periodic kernel if both belonged to the periodic function, and zero otherwise. The prior mean in the change-point composition was set to the means learned in linear and periodic sub-tasks respectively. The model acted using an ϵ-greedy policy with an ϵ-value of 0.9 in the first two sub-tasks and acted greedily on the last sub-task. These ϵ-values were chosen such that the model explored and learned the latent functions in the first two sub-tasks and exploited the learned latent functions in the last sub-task.

Meta-Reinforcement Learning: The prior over possible reward functions and how these are composed are hard-coded in the Gaussian Process model. To test whether such priors and compositions can also be learned from experience, we implemented a meta-reinforcement learning agent similar to that of [5]. The agent architecture consisted of a long short-term memory (LSTM) network followed by an actor-critic module. The LSTM network had 48 hidden units and takes the trial index, the action from the previous step, the reward from the previous step, and – depending on the condition – the sub-task index as well as a one-hot encoding of the compositional rule as inputs. The actor-critic module comprised of a two-layer neural network with 48 hidden units in each layer with the first layer shared between actor and critic. It outputted an estimate of the value function and the policy. We trained the model using the standard A2C loss at the end of each episode with an additional entropy regularization term to prevent premature convergence [6]. The parameter that controlled the strength of the entropy regularization term was annealed to zero over the course of training. We used the ADAM optimizer with a learning rate of 0.001 and trained all agents for a total of $2 \cdot 10^5$ episodes.



(a) Models: Additive (b) Models: Change-Point (c) Human: Additive (d) Human: Change-Point

Figure 3: **Model and Human Performance**: (a-b) Mean regrets for both models averaged over 100 runs in the final sub-task for the additive and change-point rule. (c-d) Mean regrets for participants in the final sub-task for the additive and change-point rule. Error-bars for models show standard errors over five random seeds; for human data, standard errors were computed over participants.

2.3 Model Simulations

We simulated both of these models on the task described in Section 2.1. To investigate whether the models are able to reason compositionally, we measured their mean regrets in the final sub-task. The regret is defined as the difference between the optimal reward for a given latent function and the reward of the action selected by the agent. The result of this analysis is shown in Figure 3 (a) and (b). For both models, we find that the initial mean regret in the curriculum condition is close to zero. This indicates that the agents re-used earlier learned functions to compose new functions in a zero-shot manner. In contrast to this, the mean regret in the non-curriculum condition starts at chance-level and takes three to four trials to reach similar performance.

2.4 Human Experiment

How does human behavior compare to these two algorithms? To answer this question, we conducted a behavioral study in which the experimental design followed the previously outlined task structure. Participants were told that they are gamblers visiting the fictional town of Bandit City. This entailed visiting multiple casinos in order to play different sets of slot machines. Each casino had two slot machines by different companies, with their color indicating the manufacturer. Coins were earned after choosing an option on the slot machine. Participants were told that all slot machines from a company behaved similarly. That is, the expected payoffs across options for machines from the same company followed a similar function (either linear or periodic) but were provided no information about which function belonged to which company.

In each casino, participants had five trials per slot machine, and their goal was to win as many coins as possible. Afterward, they were then tested on a new slot machine, which was a composition of the two previously played machines. The rewards of this new slot machine were given by either an additive or a change-point composition. Machines with an additive composition had options that were half-green and half-blue in color, while options on machines with a change-point composition were either green or blue depending on the machine from which its reward was drawn (as shown in Figure 2).

Participants: We recruited 200 participants (103 female, M_{age} = 28.90) for the additive and 211 participants (96 female, M_{age} = 27.58) for the change-point composition through the Prolific platform. Participants were randomly assigned to the curriculum or non-curriculum condition. Each condition involved 20 tasks in total. 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 $\pounds 2$ and a performance-dependent bonus payment up to $\pounds 2.5$. The study was approved by the local ethics committee. We removed 25 participants from our analysis as their performance on the last sub-task was not above chance level.

Results: If people can learn to compose in our task, we would expect them to pick the most rewarding option in the curriculum condition on the first trial of the final sub-task with a higher probability than people in the non-curriculum condition. We consider the last task as an evaluation task as we are interested in the converged behavior. The main quantity of interest is again the regret in the compositional sub-task. Figure 3 (c) and (d) shows the mean regret (averaged over participants) for both compositional rules.

A mixed linear regression analysis revealed that participants in the curriculum condition had a significantly lower regret on the first trial of the final sub-task compared to participants in the non-curriculum condition ($\hat{\beta} = -1.49 \pm 0.06, p < .001$). This result indicates that people were able to perform some form of zero-shot compositional reasoning in our task, similar to the investigated models.

However, there were also significant differences to the behavior of both models. The regret in the compositional sub-task for the participants was larger than that of the Bayesian and the meta-learned model. This is to be expected, as both models provide us with idealized predictions. Participants furthermore kept on learning during the final sub-task in the curriculum condition, while both computational models did not. To quantify this effect we fitted a mixed-effects linear regression using per-trial regret 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 ($\hat{\beta} = -0.31 \pm 0.02, p < .001$) onto regret. This confirms that the performance of participants improved with additional interactions. The improvement in the curriculum condition ($\hat{\beta} = -0.21 \pm 0.02$) however was weaker than that in the non-curriculum condition ($\hat{\beta} = -0.42 \pm 0.02$), indicating that participants could exploit the given task structure.

3 Discussion

People effortlessly compose previously learned functions into new ones when navigating their everyday lives. Previous research has examined this ability in both humans and artificial agents [7, 8] but in a different setting. In the present article, we proposed a novel experimental task to probe how different agents generalize compositionally. We found that two very different computational models – a Gaussian Process model and a meta-learned recurrent neural network – can compose previously learned functions into a new one in a zero-shot manner. We furthermore compared the behavior of these models to that of human subjects. While we demonstrated that people can learn to compose reward functions, our analysis also revealed a gap to the behavior of both models. We hope to address this gap in future work by building agents that take the complexity of their solutions into account [9].

It is generally thought that Bayesian models and neural networks offer different theories for understanding human cognition. In neural network models, cognition emerges from the interaction between a large set of simple processing units. Bayesian models, on the other hand, explain cognition by appealing to the idea of ideal statistical inference. The concept of meta-learning offers a bridge between the two frameworks. In the current study, we have seen that it is possible to meta-learn a recurrent neural network that can reason compositionally on a challenging task – an ability that has been historically investigated using Bayesian models. In general, we believe that future work should view both frameworks not as competing hypotheses but rather as symbiotic tools. Each of them comes with its own advantages and disadvantages: Bayesian models are often interpretable but hard to scale, while meta-learning offers an answer to the question of how to learn useful priors from experience but comes at the cost of losing theoretical guarantees.

References

[1] David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning*, 2013.

[2] Eric Schulz, Nicholas T Franklin, and Samuel J Gershman. Finding structure in multi-armed bandits. *Cogn. Psychol.*, 119:101261, 2020.

[3] Daniel E Acuña and Paul Schrater. Structure learning in human sequential decision-making. *PLoS Comput. Biol.*, 6(12):e1001003, 2010.

[4] Tankred Saanum, Eric Schulz, and Maarten Speekenbrink. Compositional generalization in multi-armed bandits. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

[5] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. In *CogSci*, 2017.

[6] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

[7] Brenden M Lake. Compositional generalization through meta sequence-to-sequence learning. *Advances in neural information processing systems*, 32, 2019.

[8] Sreejan Kumar, Ishita Dasgupta, Jonathan Cohen, Nathaniel Daw, and Thomas Griffiths. Meta-learning of structured task distributions in humans and machines. In *International Conference on Learning Representations*, 2020.

[9] Marcel Binz, Samuel J Gershman, Eric Schulz, and Dominik Endres. Heuristics from bounded meta-learned inference. *Psychological review*, 2022.