A Generative Grammar for Automatically Designing Experiments on Human Learning and Decision Making

Maria Eckstein (mariaeckstein@deepmind.com)
Google DeepMind

Kevin Miller (kevinmiller@google.com)
Google DeepMind and University College London

Angela Radulescu (angela.radulescu@mssm.edu)
Center for Computational Psychiatry, Icahn School of Medicine at Mount Sinai

Abstract

The study of human learning and decision making has fascinated many researchers for a long time. As a result, the number of experimental paradigms in this field is large, and increasingly more sophisticated experiments are added continuously. While this multitude of approaches has provided innumerable insights, the resulting fragmentation has also led to wide-ranging contradiction in results, which have been difficult to resolve. We propose a method that leverages the strength of using multiple tasks to study a complex phenomena, while mitigating its disadvantages. Our method involves the specification of a family of tasks, defined by a procedural grammar that is based on a small number of task features. Rather than designing tasks individually, the grammar allows sampling them across the allowed space defined by the features. A dataset of human choices collected using this method is expected to reveal foundational insights about human learning and decision making that are generalizable and robust to task variations.

Keywords: reinforcement learning; procedurally-generated tasks; meta-learning

Introduction

In humans, learning is remarkably general: we are able to adapt our cognitive strategies to achieve good performance in a broad range of situations. Similarly, a holistic theory of human learning should account for behavior across a broad diversity of tasks. A typical research study investigating human learning, however, often considers only one task, as it is designed to answer one particular question. This approach has two important drawbacks. Firstly, even simple tasks are known to elicit a variety of complex cognitive processes, which may differ dramatically from those that the task was designed to study (Collins & Frank, 2012; Radulescu, Vong, & Gureckis, 2022; Wimmer, Braun, Daw, & Shohamy, 2014). Secondly, even seemingly-related tasks can elicit strikingly different behavior (Eckstein, Master, Xia, et al., 2022; Nussenbaum & Hartley, 2019). Together, this means that it is often difficult to predict how results obtained using one task will generalize to another, and even more difficult to assemble results obtained using multiple tasks into a coherent theory of human learning.

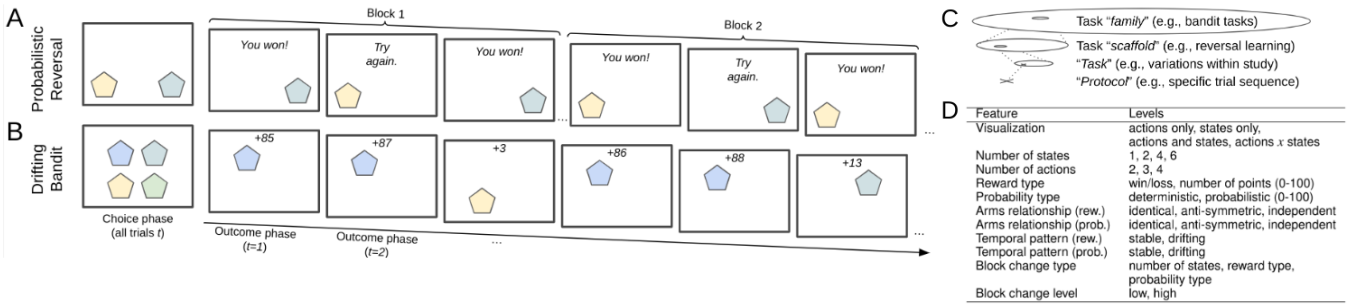
Other fields have seen progress on these issues by considering a large number of tasks (for similar approaches, see, e.g., (Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021; Almaatouq et al., 2024)). Here, we propose to use a generative "grammar" to define a desired range of tasks in the framework of human learning and decision making. This grammar can be used to express context-free bandit tasks, in which subjects select on each trial one of a set of discrete available actions, receive a reward outcome, and are tasked with optimizing total reward received over the course of the task. Each production of this grammar is a particular task protocol that defines everything that is needed to run an experiment. The nature of the grammar makes explicit the structured relationships between various tasks.

A Generative Grammar for Bandit Tasks

In constructing our grammar, we found it helpful to think about experimental datasets as a layering of levels (Fig. 1C): At the bottom is the experimental "protocol": the specific sequence of trials experienced by one participant in one study. Next is the "task": the combination of all protocols in a dataset; oftentimes, there are small variations between protocols in a dataset, e.g., different random seeds for different participants. Next is what we call the task "scaffold": a collection of tasks that are aimed to answer similar research questions, and that share specific design choices that make them distinct from other scaffolds. For example, the scaffold of probabilistic reversal tasks has been developed to study how humans adapt to switches in their environment; here, reward contingencies (e.g., which of two action wins versus loses) switch unpredictably over time (e.g., (Cools et al., 2009; Eckstein, Master, Dahl, Wilbrecht, & Collins, 2022), Fig. 1A). A different scaffold, multi-armed drifting bandit tasks, has been used to study how humans learn to decide between multiple options with fluctuating payoffs (e.g., quality of different restaurants over time) (e.g., (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006), Fig. 1B). At the top of the layering is the task "family", the creation of which is the goal of the current study. In the study of learning and decision making, this family includes both of the scaffolds mentioned above, as well as most bandit task.

At the heart of our method is the concept of a *task grammar*, which defines the *family of tasks* within a paradigm. The grammar specifies a vocabulary of task features (Fig. 1D) and how these features can be used to construct task scaffolds, from which individual protocol are then derived. We created a grammar that encompasses a large fraction of existing learning and decision making tasks, aiming to provide good coverage of this research paradigm. Tasks sampled from our grammar include both existing task scaffolds and new scaffolds that combine existing features in different ways.

The grammar's vocabulary specifies the *task features* that all tasks in a family can vary on. Our goal was to distill the largest number of existing tasks into the smallest number of task dimensions, each with the smallest number of features, to obtain a grammar that is *maximally expressive* (able to create as many existing and interesting new task paradigms as possible); and at the same time *maximally constrained*, i.e. providing enough scaffold to strongly favor "meaningful" tasks. We will explain the task features using our two example scaffolds, probabilistic reversal (PR; Fig. 1A) and multi-armed drifting bandit (MDB; Fig. 1B). The PR and MDB scaffolds differ on 6 task features out of the 11 we identified (Fig. 1D): the number of choice options (2 for PR vs. 4 for MDB), the reward type (win/loss for PR vs. number of points for MDB), the probability type (probabilistic outcomes for PR vs. deterministic outcomes for MDB), the relationships between arms in terms of rewards (identical outcomes for both arms for PR vs. independently sampled outcomes for MDB), the relationship between arms in terms of probabilities (anti-correlated outcome probabilities for PR vs. identical, deterministic out-



Feature	Levels
Visualization	actions only, states only, actions and states, actions x states
Number of states	1, 2, 4, 6
Number of actions	2, 3, 4
Reward type	win/loss, number of points (0-100)
Probability type	deterministic, probabilistic (0-100)
Arms relationship (rew.)	identical, anti-symmetric, independent
Arms relationship (prob.)	identical, anti-symmetric, independent
Temporal pattern (rew.)	stable, drifting
Temporal pattern (prob.)	stable, drifting
Block change type	number of states, reward type, probability type
Block change level	low, high

Figure 1: Example trial sequences of A) probabilistic reversal task and B) 4-armed drifting bandit task. C) Hierarchy of datasets. D) Task features used to define the grammar.

comes for MDB), and finally, the temporal pattern of rewards (stable within blocks for PR vs. drifting for MDB). Completing the list of 11, both scaffolds have the same temporal pattern of probabilities (stable within block), the same visualization (only actions are visible), and the same number of "states" (1; where each state defines one particular action-outcome mapping). Furthermore, PR requires specific changes to occur at block boundaries: the block change needs to be of type "reward type" and of level "low", to enable the resampling of outcome probabilities (further details about the features are omitted for brevity).

While PR and MDB have been used to study distinct research questions, expressing both using the same features reveals that we can interpolate between them, creating new scaffolds that inherit some features from one, and some from the other. Our method can be used to automatize this process by constructing a task grammar based on the 6 combined features of these 2 tasks, resulting in a family of $2^6 = 64$ distinct scaffolds. What could a new scaffold in this space look like? The grammar might sample a 4-armed, binary (win/loss), probabilistic, blocked task scaffold with anti-correlated outcome probabilities. Upon closer inspection, such a paradigm would be crucial to assess how people perform complex, multi-dimensional state inference: When some options are good, specific others are bad, and vice versa; furthermore, changes in the contingencies of one option imply changes for all others (e.g., if strawberries and peaches are in season, turnips and potatoes are not, and vice versa). Each of the remaining 61 novel task scaffolds defined by this grammar might be equally relevant for our understanding of human learning and decision making.

Our final vocabulary (Fig. 1D) specifies the features that are required to recreate a wide range of existing learning and decision making tasks, e.g., (Cools et al., 2009; Eckstein, Master, Dahl, et al., 2022; Collins & Frank, 2012; Daw et al., 2006; Behrens, Woolrich, Walton, & Rushworth, 2007; Palminteri, Khamassi, Joffily, & Coricelli, 2015; Davidow, Foerde, Galvan, & Shohamy, 2016; Frank et al., 2015; Collins & Koechlin, 2012; Frank, Seeberger, & O'Reilly, 2004; Klein, Ullsperger, & Jocham, 2017) and many others. Comprising 11 features with 2-4 levels each, the task family defines $4*4*3*2*2*2*2*3*3*3*2 = 41,472$ unique task scaffolds, of which only a few dozens have been studied so far.

The grammar works end-to-end. On each iteration, it randomly creates one scaffold, which is the blueprint for a near-infinite number of possible tasks. One of them will be sampled at random and turned it into one of a near-infinite number of possible protocols, which can by used without further processing by our experimental software to collect participant data.

Discussion

The current learning and decision making literature is showing increasing fragmentation, both in terms of tasks and in terms of cognitive models used to analyze them. We propose a method that streamlines the creation of new tasks based on a desired range of features, in a hope to counteract negative consequences of fragmentation, while leveraging the potential of using multiple tasks to study a complex cognitive phenomenon. "Filling in the gaps" between existing tasks will shed additional light on the cognitive relevance of long-standing concepts (e.g., stochasticity, volatility), and potentially pave the way towards a more comprehensive understanding of learning and decision making as a whole.

We want to note that our grammar is not complete. E.g., we have not (yet) included feature-based bandits, varying time horizons, visibility of counter-factual outcomes, multi-step tasks, or prediction tasks. However, it is easy to add new features to the grammar and adapt it as needed. We foresee a trade-off in which a grammar with more features allows coverage of a richer task space, but at the cost of exponentially larger gaps between tasks, due to the exponential explosion of the family size with increasing numbers of features. Empirical work will be required to determine the best specification of the grammar for each research question. We will collect five pilot datasets; four will be based on the full grammar, one at each level of abstraction; a fifth will be based on a subgrammar defined by just two tasks (PR and MDB). We will use flexible, neural-network based methods (Eckstein, Summerfield, Daw, & Miller, 2023; Miller, Eckstein, Botvinick, & Kurth-Nelson, 2023) to analyze these datasets, and inform future development of the method.

References

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2024, January). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, *47*, e33. doi: 10.1017/S0140525X22002874

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007, September). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. doi: 10.1038/nn1954

Collins, A. G. E., & Frank, M. J. (2012, April). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis: Working memory in reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1024–1035. doi: 10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Koechlin, E. (2012). Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLOS Biology*, *10*(3), e1001293. doi: 10.1371/journal.pbio.1001293

Cools, R., Frank, M. J., Gibbs, S. E., Miyakawa, A., Jagust, W., & D'Esposito, M. (2009, February). Striatal Dopamine Predicts Outcome-Specific Reversal Learning and Its Sensitivity to Dopaminergic Drug Administration. *Journal of Neuroscience*, *29*(5), 1538–1543. doi: 10.1523/JNEUROSCI.4467-08.2009

Davidow, J., Foerde, K., Galvan, A., & Shohamy, D. (2016, October). An Upside to Reward Sensitivity: The Hippocampus Supports Enhanced Reinforcement Learning in Adolescence. *Neuron*, *92*(1), 93–99. doi: 10.1016/j.neuron.2016.08.031

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006, June). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. (Number: 7095 Publisher: Nature Publishing Group) doi: 10.1038/nature04766

Eckstein, M. K., Master, S. L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. E. (2022, April). Reinforcement learning and bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal. *Developmental Cognitive Neuroscience*, 101106. doi: 10.1016/j.dcn.2022.101106

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. (2022, November). The interpretation of computational model parameters depends on the context. *eLife*, *11*, e75474. doi: 10.7554/eLife.75474

Eckstein, M. K., Summerfield, C., Daw, N. D., & Miller, K. J. (2023, May). *Predictive and Interpretable: Combining Artificial Neural Networks and Classic Cognitive Models to Understand Human Learning and Decision Making.* bioRxiv. (Pages: 2023.05.17.541226 Section: New Results) doi: 10.1101/2023.05.17.541226

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015, January). fMRI and EEG Predictors of Dynamic Decision Parameters during Human Reinforcement Learning. *Journal of Neuroscience*, *35*(2), 485–494. (Publisher: Society for Neuroscience Section: Articles) doi: 10.1523/JNEUROSCI.2036-14.2015

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science*, *306*(5703), 1940–1943. doi: 10.1126/science.1102941

Klein, T. A., Ullsperger, M., & Jocham, G. (2017, June). Learning relative values in the striatum induces violations of normative decision making. *Nature Communications*, *8*(1), 16033. (Publisher: Nature Publishing Group) doi: 10.1038/ncomms16033

Miller, K., Eckstein, M., Botvinick, M., & Kurth-Nelson, Z. (2023, December). Cognitive Model Discovery via Disentangled RNNs. *Advances in Neural Information Processing Systems*, *36*, 61377–61394.

Nussenbaum, K., & Hartley, C. A. (2019, December). Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental Cognitive Neuroscience*, *40*, 100733. doi: 10.1016/j.dcn.2019.100733

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015, August). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*(1), 8096. (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/ncomms9096

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021, June). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214. doi: 10.1126/science.abe2629

Radulescu, A., Vong, W. K., & Gureckis, T. M. (2022). Name that state: How language affects human reinforcement learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Wimmer, G. E., Braun, E. K., Daw, N. D., & Shohamy, D. (2014, November). Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(45), 14901–14912. doi: 10.1523/JNEUROSCI.0204-14.2014