Building Deterministic Causal Graphs Using Reinforcement Learning in Cognitive Tasks with Evolutionary Bias

Ines Aitsahalia (ifa2108@cumc.columbia.edu)
Center for Theoretical Neuroscience, Columbia University, 3227 Broadway
New York, NY 10027, USA

Adithya Gungi (ag4472@columbia.edu)
Department of Physics, Columbia University 538 West 120th Street
New York, NY 10027, USA

Pradyumna Sepúlveda (ps3345@columbia.edu)
Department of Psychiatry, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center
New York, NY 10032, USA

Kiyohito Iigaya (ki2151@columbia.edu)
Center for Theoretical Neuroscience, Columbia University, 3227 Broadway
New York, NY 10027, USA

Abstract

Despite the importance of uncovering causal, rather than correlational, structures in the world to survival, algorithms for this type of causal learning remain computationally taxing. Recent neural evidence has challenged the ability of reinforcement learning (RL) algorithms to provide a useful approximation. Here, we present a new reinforcement learning model that uses modified successor representations and incorporates evolutionary death avoidance, capturing a wide variety of human structure learning and animal conditioning. To formally capture the risk of learning in the wild we implement a constraint where punishment distributions are inherently heavy-tailed to account for the risk of death. This places the intrinsic value on having a deterministic graph in this framework, parsimoniously capturing a wide range of instrumental and non-instrumental behaviors.

Keywords: reinforcement learning; causal learning; intervention; structure learning; death; risk avoidance; evolution

Introduction

Learning in the wild is dangerous. Animals constantly face the possibility of death if they take wrong actions. This, death, is a largely overlooked aspect in standard reinforcement learning models, where agents are often allowed to learn through trial and error (death) to collect more rewards. However, in reality, there is no trial after death. How do animals solve this problem?

One way to avoid death is to discover and exploit causal, rather than correlational, structures in the world. With causal structural knowledge, animals can avert actions leading to death and exploit the safe structures that will lead to rewards. However, discovering such causal structures often requires animals to explore and *intervene* in the world, which can come with a risk of death. This suggests an inherent tradeoff between exploration and exploitation in causal learning. How do animals address this tradeoff and learn the causal structures of the world? Recent behavioral and neural evidence (Jeong et al., *Science* 2022) has challenged the sufficiency of reinforcement learning (RL) models in capturing causal learning and its underlying neural mechanisms. Here, we propose a novel computational model that captures causal structure learning, while at the same time preserving the RL paradigm with an evolutionary bias for death avoidance.

Existing model-based computational frameworks, such as model-based RL, typically assume predefined world structures and avoid structure learning. Bayesian inference models, although capable of learning complete probabilistic structures, lack biological plausibility due to their high computational demands. Most existing models also overlook the human inclination to infer deterministic structures even when none exist (Redelmeier and Tversky, 1996).

To address these gaps, we propose a computational model that builds a *deterministic causal graph* based on both observational and interventional learning. Our model leverages a nonlinear transformation of the successor representation (Dayan, 1993), learned through a basic RL algorithm. Our model explains various empirical findings, including spatial navigation, classical conditioning, a mixture of model-based vs. model-free learning, and information-seeking behavior that standard RL models cannot capture.

Taken together, our work offers a novel method to bridge RL and causal learning in a biologically and cognitively constrained way. Our findings suggest the deterministic causal graph as a unifying mechanism for a range of phenomena studied separately.

Algorithm details We introduce a variant of the successor representation (SR), which we refer to as transitional successor representation (tSR). Unlike the standard SR that learns discounted future *occupancy*, the tSR directly learns discounted future *transition rates* at each time step t in a

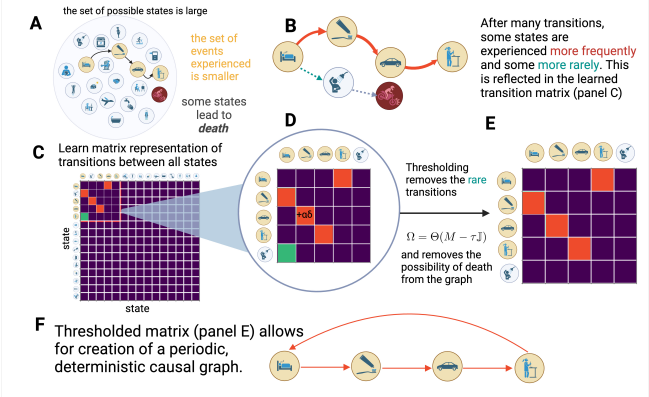

Figure 1: (**A**) The set of all possible experiences discretized into states is very large, with some states leading to death. (**B**) The sequences observed in life are smaller, and the transition probabilities between states are important. (**C**) Learning the transition successor representation (tSR). (**D**). Thresholding the tSR to create a binary matrix. (**E**). Constructing the deterministic causal graph (dCG) using the thresholded matrix. **F**)The learned causal graph.

multi-step transition matrix, using a simple temporal difference learning rule (**Fig. 1C**):

$$M(s_t, s') \rightarrow M(s_t, s') + \alpha\delta^{\mathrm{SR}}_{s_t}(s') \tag{1}$$

where $M(s_t, s')$ is the tSR matrix describing transitions from state s_t to state s', α is the learning rate, and $\delta^{\mathrm{SR}}_{s_t}(s')$ is the temporal difference error:

$$\delta^{\mathrm{SR}}_{s_t}(s') = I\left(s_{t+1} = s'\right) + \gamma M\left(s_{t+1}, s'\right) - M\left(s_t, s'\right), \tag{2}$$

with I the identity matrix, and γ the discounting rate. To obtain a **deterministic causal graph** (dCG), Ω, we apply a simple threshold to the probabilistic tSR matrix M (**Fig. 1D**):

$$\Omega(s_t, s') = \Theta\left(M(s_t, s') - \tau\right), \tag{3}$$

where Θ denotes the Heaviside step function and τ is the threshold parameter. Through this nonlinear operation, the model approximates probabilistic world structures by functionally binarizing them, creating a compact causal graph. This not only captures the human tendency to perceive deterministic structures even in the absence, but also makes the algorithm efficient and biologically plausible.

Evolutionary heavy-tailed bias We implement an evolutionary bias into the successor representation, an assumption that builds in a small probability that any states will eventually lead to death, as a biological constraint in the model, where death has an infinitely large negative value. This potential death always exists in the states that have yet to form a causal graph. Creating the causal graph eliminates the possibility of death from states within the graph, providing an intrinsic, infinitely large, bonus to discovering deterministic graphs. Due to the possibility of death, punishment is inherently heavy-tailed, providing a critical insight to distributional RL. Importantly, in this algorithm, the agent's actions are determined by the causal graph, Ω, instead of tSR, M.

Thus our model makes distinct predictions from classic SR in RL frameworks, boosted by the value of avoiding death. For instance, the model predicts that the agent first explores to discover a causal graph. Once the agent identifies the graph that safely leads to a reward, the agent will exploit the graph repeatedly. This explore-exploit behavior is consistent with observed non-instrumental information-seeking. Constructing the model with multiple discount factors γ (like in the distributional RL; Masset et al., 2023) enables the agent to learn causal structures across multiple timescales.

Results

Our causal graph model captures the well-studied mixture of model-based and model-free learning in the two-step tasks (Daw et al., *Neuron* 2011) (**Fig. 2A**) without resorting to the dichotomy (**Fig. 2B**). Our model also captures a wide range of empirical findings that standard RL models struggle to account for, such as human structure learning (Momennejad et al., 2017), blocking and unblocking in classical conditioning (Maes et al., 2018), and information seeking behavior (Bromberg-Martin and Hikosaka, 2009); as well as recent key findings on causal learning (Jeong et al., 2022) in animals.

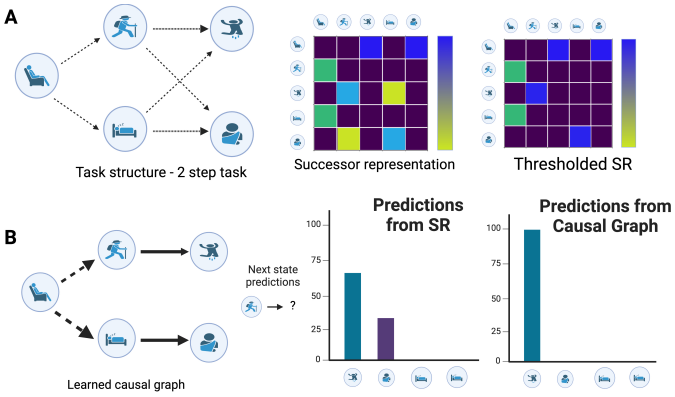

Figure 2: Predictions of the model on 2-step task. (**A**). Structure of the 2-step task. The original tSR for this task is the same, however the threshold eliminates the rare transitions. (**B**) The threshold creates a dCG that influences behavior, making predictions that can capture both model-based and model-free learning by tuning the threshold.

As we describe for the 2-step learning task (**Fig. 2A**), the model agent builds a deterministic causal graph (**Fig 1F**) and generates behaviors that resemble empirical data, while the standard SR model falls short (**Fig 2B**).

Our model offers new testable predictions in complex inference and exploration tasks. For example, we can explain information-seeking (Bromberg-Martin and Hikosaka, 2009)

without specifically bonusing new information, but rather by allowing the inherent value of determinism to boost information-seeking behaviors. We also predict an increased preference for information on a longer timescale (as tested in Iigaya 2020). We are currently piloting a novel behavioral task aimed at learning from direct causal interventions.

In sum, our study offers a novel computational framework to understand causal learning, using an RL algorithm to construct deterministic causal graphs. By incorporating death as an avoidable consequence of state transitions, our model explains a wide range of behaviors studied separately as a consequence of common causal structure learning. Our findings highlight how heavy-tailed punishment distribution with death may have shaped a range of our instrumental and non-instrumental behavior.

.

Acknowledgments

We thank Kim Stachenfeld, Ken Miller, Sashank Pisupati, Isabel Berwian, and Rebecca Lysaght. This work is supported by the BBRF NARSAD Young Investigator Grant (KI), the Saks Fifth Avenue Transformational Depression Research Award (KI), and NSF GRFP Fellow ID: 2022342390 (IA).

References

Bromberg-Martin, E. S., & Hikosaka, O. (2009, July). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–126. Retrieved from http://dx.doi.org/10.1016/j.neuron.2009.06.009 doi: 10.1016/j.neuron.2009.06.009

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011, March). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. Retrieved from http://dx.doi.org/10.1016/j.neuron.2011.02.027 doi: 10.1016/j.neuron.2011.02.027

Dayan, P. (1993, July). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624. Retrieved from http://dx.doi.org/10.1162/neco.1993.5.4.613 doi: 10.1162/neco.1993.5.4.613

Iigaya, K., Hauser, T. U., Kurth-Nelson, Z., O'Doherty, J. P., Dayan, P., & Dolan, R. J. (2020, June). The value of what's to come: Neural mechanisms coupling prediction error and the utility of anticipation. *Science Advances*, *6*(25). Retrieved from http://dx.doi.org/10.1126/sciadv.aba3828 doi: 10.1126/sciadv.aba3828

Jeong, H., Taylor, A., Floeder, J. R., Lohmann, M., Mihalas, S., Wu, B., ... Namboodiri, V. M. K. (2022, December). Mesolimbic dopamine release conveys causal associations. *Science*, *378*(6626). Retrieved from http://dx.doi.org/10.1126/science.abq6740 doi: 10.1126/science.abq6740

Maes, E., Krypotos, A.-M., Boddez, Y., Alfei Palloni, J. M., D'Hooge, R., De Houwer, J., & Beckers, T. (2018, April). Failures to replicate blocking are surprising and informative—reply to soto (2018). *Journal of Experimental Psychology: General*, *147*(4), 603–610. Retrieved from http://dx.doi.org/10.1037/xge0000413 doi: 10.1037/xge0000413

Momennejad, I. (2020, April). Learning structures: Predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences*, *32*, 155–166. Retrieved from http://dx.doi.org/10.1016/j.cobeha.2020.02.017 doi: 10.1016/j.cobeha.2020.02.017

Redelmeier, D. A., & Tversky, A. (1996, April). On the belief that arthritis pain is related to the weather. *Proceedings of the National Academy of Sciences*, *93*(7), 2895–2896. Retrieved from http://dx.doi.org/10.1073/pnas.93.7.2895 doi: 10.1073/pnas.93.7.2895