Testing causal hypotheses through Hierarchical RL

Anonymous Author(s)
Affiliation
Address
email

One goal of AI research is to develop agentic systems capable of operating in open-ended environments with the autonomy and adaptability akin to a scientist in the world of research. An ideal "AI scientist" should be able to generate and test hypotheses, and draw conclusions about the world based on the evidence. It also needs to be intrinsically motivated to adapt to a continually changing world with sparse reward signals. Here, we propose hierarchical reinforcement learning (HRL) as a key ingredient to building agents that can systematically generate and test hypotheses that enables transferrable learning of the world, and discuss potential implementation strategies.

Defining hypothesis. For us, a hypothesis is a fundamentally a statement about the causal structure of the world, which we formulate as a Structural Causal Model (SCM, Pearl (2000)). The learning objective is identifying the set of nodes (concepts) and edges (relationships) in the SCM. The focus on causality is crucial for two main reasons. First, having the right causal structure allows the agent to adapt more quickly in the face of changing environments (Bengio et al., 2019). Second, causal structures can enable the agent to more efficiently achieve its objectives via counterfactual reasoning and long-term credit assignment (Meulemans et al., 2023).

Hypothesis testing through HRL. We choose the RL framework due to its emphasis on active learning and natural interpretation of actions as interventions, and propose one way to combine Markov Decision Processes (MDPs) with SCMs (see Appendix A). Hypothesis testing through HRL leverages learned abstract-level subgoals, such as skills (Eysenbach et al., 2019) or options (Sutton et al., 1999; Bacon et al., 2016), to intervene on SCM nodes. This approach can be implemented by training hypothesis-conditioned policies, $\pi(a|s, h)$, where the hypothesis h consists of variables with different attributes. For example, in the blicket detector task from developmental psychology, we can formulate hypotheses about relationships between variables representing objects and the detector's outcome (Gopnik and Sobel, 2000). Consider a scenario with three potential blickets ($X^{(1)}, X^{(2)}, X^{(3)}$) and a blicket machine ($X^{(4)}$). A hypothesis might be that $X^{(1)} = \texttt{on\_top\_machine}$ leads to $X^{(4)} = \texttt{on}$, indicating that the first object is the blicket. To test this hypothesis, we would set $X^{(1)}$ to have the attribute `on_top_machine` and observe the resulting state of $X^{(4)}$, while also verifying that this relationship holds regardless of the attributes of other variables. As the action space may not directly correspond to causal interventions, we require sequences of actions (i.e. hypothesis-conditioned policies) to set variables to specific attributes, therefore allowing the agent to observe the outcome of interventions. This naturally gives rise to an HRL setting where action sequences occur at lower temporal abstractions than the world model reasoning about variable relationships. Further, our HRL approach is also inspired by cognitive science, particularly the observation that humans act and plan at abstract rather than muscle level, and that children are "scientists in the cribs" (Gopnik et al., 2009) who excel at learning efficiently the causal structure of the world through exploration and experimentation. Hypothesis testing, through this lens, could be seen as a way of guiding exploration at the abstract (i.e., SCM) level. It can also be easily combined with other child-inspired intrinsic motivations, such as empowerment (Gopnik, 2024) as a way of deciding which hypothesis to test (see Appendix B).

In conclusion, here we present a framework for designing AI agents that can generate and test hypotheses using HRL, inspired by developmental psychology, and propose some concrete implementations. We hope to prompt discussion about future directions, including a formal definition of hypothesis and hypothesis testing, and foster collaborations among disciplines.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

References

Bacon, P., Harb, J., and Precup, D. (2016). The option-critic architecture. CoRR, abs/1609.05140.

Bengio, Y., Deleu, T., Rahaman, N., Ke, N. R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. J. (2019). A meta-transfer objective for learning to disentangle causal mechanisms. CoRR, abs/1901.10912.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., Hughes, E., Battaglia, P., Botvinick, M., and Kurth-Nelson, Z. (2019). Causal reasoning from meta-reinforcement learning. arXiv preprint arXiv:1901.08162.

Eberhardt, F. (2007). Causation and intervention. Unpublished doctoral dissertation, Carnegie Mellon University, 93.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. (2019). Diversity is all you need: Learning skills without a reward function. In International Conference on Learning Representations (ICLR).

Gopnik, A. (2024). Empowerment as causal learning, causal learning as empowerment: A bridge between bayesian causal hypothesis testing and reinforcement learning. In PhilSci-Archive.

Gopnik, A., Meltzoff, A., and Kuhl, P. (2009). The Scientist in the Crib: What Early Learning Tells Us About the Mind. HarperCollins.

Gopnik, A. and Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. Child development, 71(5):1205–1222.

Klyubin, A., Polani, D., and Nehaniv, C. (2005). Empowerment: A Universal Agent-Centric Measure of Control. In 2005 IEEE Congress on Evolutionary Computation, volume 1, pages 128–135, Edinburgh, Scotland, UK. IEEE.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2008). Keep Your Options Open: An Information-Based Driving Principle for Sensorimotor Systems. PLOS ONE, 3(12):e4018. Publisher: Public Library of Science.

Marino, K., Fergus, R., Szlam, A., and Gupta, A. (2020). Empirically verifying hypotheses using reinforcement learning. arXiv preprint arXiv:2006.15762.

Meulemans, A., Schug, S., Kobayashi, S., Daw, N., and Wayne, G. (2023). Would i have gotten that reward? long-term credit assignment by counterfactual contribution analysis.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press.

Peters, J., Janzing, D., and Schölkopf, B. (2017). Elements of causal inference: foundations and learning algorithms. The MIT Press.

Salge, C., Glackin, C., and Polani, D. (2012). Approximation of empowerment in the continuous domain. Advances in Complex Systems. Accepted: 2013-01-15T14:58:59Z.

Strehl, A. L. and Littman, M. L. (2008). An analysis of model-based interval estimation for markov decision processes. Journal of Computer and System Sciences, 74(8):1309–1331.

Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. Artificial intelligence, 112(1-2):181–211.

Appendix

A MDPs and SCMs

(a) “Timeless” SCM

(b) Time dependent (Markov) SCM

Figure 1: Structural Causal Models (SCMs) describing causal relationships. The inter-variable relationships are the same between the two SCMs, with the time-dependent SCM treating the dependency as occurring across a single time-step.

We propose one perspective to reason about Markov Decision Processes (MDPs) and Structural Causal Models (SCMs) *together*. The former is a framework for embodied behavior, while the latter reasons about structures and relationships.

A (reward free) Markov Decision Process (MDP) is the tuple $\langle \mathcal{S}, \mathcal{A}, P \rangle$, with state space $\mathcal{S}$, primitive action space $\mathcal{A}$, and transition probability function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. A Structural Causal Model (SCM) is defined via a set of internal variables $X = \{X^{(1)}, X^{(2)}, ..., X^{(n)}\}$, and independent noise variables $\{\epsilon^{(1)}, \epsilon^{(2)}, ..., \epsilon^{(n)}\}$. A SCM consists of a collection of n assignments,

$$X^{(i)} \leftarrow f_i(\text{Pa}(X^{(i)}), \epsilon^{(i)}) \,, \tag{1}$$

where $\text{Pa}(X^{(i)}) \subseteq \{X^{(1)}, ..., X^{(n)}\} \setminus \{X^{(i)}\}$ are the *parents* of $X^{(i)}$, and f_i is some function that takes the parent nodes’ values as inputs to determine the child node’s value (Peters et al., 2017).

Including the Notion of Time One often reasons about an SCM as “timeless” and encoding invariant facts about the world. To reason about how variables evolve dynamically over time, we instead treat an assignment as *invariant across time-step*. Specifically, instead of considering the causal parent of $X^{(i)}$, we consider the causal parent of $X^{(i)}$ *at time* t:

$$X_t^{(i)} \leftarrow f_i(\text{Pa}(X_t^{(i)}), \epsilon^{(i)}) \,. \tag{2}$$

If we further make the Markov assumption,[1] then the variables in X_t are independent of *all* other variables given X_{t-1}. In other words, the parents of any variable $X_t^{(i)}$ must belong to the set X_{t-1}, i.e. $\text{Pa}(X_t^{(i)}) \subseteq X_{t-1}$. An example of such a *time-dependent SCM* is illustrated in Figure 1.

Actions and Interventions A common type of intervention are *structural*, or “surgical” interventions. Such interventions break (i.e. make independent) a variable $X^{(i)}$ from its causal parents and set it to a particular value (i.e. $P(X^{(j)} | \text{do}(X^{(i)} = c))$). In specific settings, actions in an MDP can correspond exactly to structural interventions (Dasgupta et al., 2019). Generally speaking, actions do not make variables fully independent of its causal parents, but only *influence* its value. This is referred to as a *parametric* intervention (and is related to the idea of instrumental variables). For a fuller discussion of the two types of interventions, we refer the reader to Eberhardt (2007).

States as Variable Sets The first way of combining together the two frameworks is simply to treat the set of structural variables *as* a state in an MDP. I.e. $\mathcal{X} = \mathcal{S}$, and $S_t = X_t = \{X_t^{(1)}, ..., X_t^{(n)}\}$. The problem of learning the correct SCM then correspond to learning how each “state factors” ($X^{(i)}$) and actions $A_t \in \mathcal{A}$ influence factors at the next time-step. This correspond to learning a “good”

[1] Whether or not the Markov assumption is a reasonable assumption here is open for discussion, nevertheless we argue it is a useful first step in bridging together MDPs and SCMs, and opens up a set of new perspective.

state-level world model: $\Pr(S_{t+1}|S_t, A_t)$. This learning process to identify how factors influence each other can be complex, and might benefit from intrinsic rewards. Intrinsic rewards can be designed to encourage exploration of different states (i.e., different combinations of variables values). For example, an intrinsic reward might be given for visiting states that are less frequently visited (Strehl and Littman, 2008).

Hierarchies and Abstract Variables To treat a low level state S_t as the set of variable X_t is somewhat unwieldy: one has to account for small fine-grained changes at a low level (e.g. modelling "as I move left for one time-step, what is the effect of this on my pixel observation of the world"). Instead, it may be much more natural to reason about structural variables at a more abstract level than the low level MDP states. Suppose we have a mapping from generic MDP states to a small set of structural variables: $\mathrm{M} : \mathcal{S} \to \mathcal{X}$. And further we can consider temporally extended *action sequences* instead of primitive actions as the interventions (Marino et al., 2020). The problem of learning the correct SCM then becomes one of learning how abstract variables and policies influence future abstract variables across multiple time-steps—an *abstracted world model*. Given $b \in \mathcal{B}$ as the set of action sequences (options / skills), we learn $\mathbf{T}(X_{u+1}|X_u, b_u)$ where the abstract time index u updates at a slower frequency than the low level time t.

By this formulation, we are not limiting ourselves to any specific state or action definition in the base MDP $\langle \mathcal{S}, \mathcal{A}, P \rangle$. Instead, through the mapping function M and the set of low level policies $\mathcal{B}$, we have define an abstracted level MDP $\langle \mathcal{X}, \mathcal{B}, \mathbf{T} \rangle$ whose states are the structural variable sets $(X^{(1)}, ...)$, and the interventions correspond to low level action sequences. The mapping function M defines what kind of concepts X_t we care about extracting from the low level states S_t, and the abstract world model learning correspond to learning the correct SCM between abstract structural variables $X_t = (X_t^{(1)}, ..., X_t^{(n)})$ and $X_{t+1} = (X_{t+1}^{(1)}, ..., X_{t+1}^{(n)})$.

B Using empowerment to select hypothesis tested

In an open-ended world with numerous potential hypotheses to test, how does one choose which to pursue for the most promising outcome? Similarly, in a scientific laboratory, what's the best approach to designing experiments that yield the most informative results? Here, we propose one potential metric of evaluating and selecting hypothesis to test: empowerment.

In the RL literature, empowerment has been used as a form of intrinsic motivation that encourages the agent to to reach situations where the agent can have more options for action, or assert greater influence on the environment (Klyubin et al., 2005). Mathematically it is defined as task-agnostic utility function via mutual information between agent's actions and outcomes: Given the random variables A (representing the sequence of K actions that the agent takes) and s' (representing the resulting states of the environment after the K actions), empowerment $\mathcal{E}$ is defined as the mutual information between A and s':

$$\mathcal{E}(A) = \mathcal{I}(A; s') = \mathbb{E}_{p(A,s')} \left[\log \left(\frac{p(A, s')}{p(A)p(s')} \right) \right]$$

Under our formulation of hypothesis as SCM, empowerment can be calculated as the mutual information between action sequence A carried out by the hypothesis-conditioned policy $\pi(a|s, h)$ and the outcome s'. One way we can choose which hypothesis to test is to select the hypothesis conditioned policies in order of their mutual information with their respective outcomes — in a way, choosing to test the hypothesis with maximal empowerment.

We note that, despite the fact that its motivation is well-rooted in cognitive science, few works have successfully deployed empowerment in an RL setting to solve real-world tasks. The main challenge is that the calculation of mutual information is computationally intractable, as it requires calculating expectations over probability distributions over s' and K-step action sequences A. This challenge is particularly significant for continuous or high-dimensional state and action spaces. Early works, such as Klyubin et al. (2005, 2008); Salge et al. (2012), stayed in discrete action spaces and used the Blahut-Arimoto algorithm, which essentially enumerates over all actions and states and thus has a high complexity. More recent works have explored the possibility of using variational inference to approximate this value. The intractability of empowerment calculation on the low level provides another strong justification for using HRL, since grouping lower-level states into abstract-level, conceptual states will reduce the number of states to iterate over, same thing for actions. Lastly, it remains an open question as to the definition of the outcome s', eg., whether it is a final state or external reward, as well as the specific implementation of estimating the mutual information.