Learning Causal Model Like Human

Zhancun Mu
YuanPei College
Peking University
yhbylch@stu.pku.edu.cn

Abstract

Humans exhibit a remarkable ability to learn causal models of their environment. This ability is crucial for understanding the world, society, and making plans and decisions. However, the current paradigm of learning and acting in an environment, primarily based on reinforcement learning (RL), lacks this causal learning ability. This leads to problems such as instability, lack of explainability, and high dependence on reward design. Most importantly, this paradigm cannot learn abstract "dark matters" like social relationships, which are mainly perceived through causal reasoning. In this essay, we will discuss features and components of humans' causal learning abilities that can provide insights into learning causal models in an environment.

1 Introduction

Building an open-ended agent with lifelong learning abilities in a real-world environment is a key goal of reinforcement learning and AI. However, this is very difficult within the current RL paradigm based on Markov Decision Process (MDP). A key limitation is the high dependence on reward design. However, reward design for diverse real-world tasks is often infeasible, including in social interactions. Another limitation is agents' high data requirements, with online interactions or offline datasets that are not transferable. This differs from humans' ability to learn from limited examples and interactions. To address these issues, one trend is world models that try to simulate predictions from agent interventions. However, this is insufficient, as we also need to understand causes and effects to guide actions and reason about other agents' intentions abstractly. This motivates the need for causal models.

Constructing causal models through exploration and experimentation remains an open challenge. To address this, we first need to understand what causality is and how humans learn it.

2 Humans' perception of causality

Causality has been widely discussed by philosophers and psychologists. One of the most famous definitions is from Hume, who reduced causality to three principles in his Regularity Theory of Causation (RTC) [12]: temporal priority, spatial contiguity, and constant conjunction. Despite a broader sense of causation existing, e.g. global warming causing sea level rise, we will focus on causation under RTC in this essay, which is sufficient. In this section, we will discuss features of humans' causal learning abilities.

2.1 Causality is a kind of belief

First, it is important to emphasize that causality is more complex than correlation. It is highly subjective and dependent on belief. Humans explore and experiment to refine beliefs about causality In fact, we establish causality not from repeated experience, but from belief even after a single occurrence. For example, when a baby sees an object placed on a music box that then starts playing, the

baby forms a strong belief that the object causes the music to play Gopnik et al. [5]. Even when the baby tries it alone and the music box does not play, the baby will still adjust the object's position. The baby updates beliefs and narrows down hypotheses in a complex way, not just through updating a correlation matrix. Why do humans sometimes strongly believe an event belongs to a causal relationship, while other times believing it is just coincidence? We argue there are differing strengths of causal beliefs.

2.2 Causal strength

How to quantify causal strength has been widely discussed. For example, Griffiths and Tenenbaum [6] uses $\Delta P = P(A|B) - P(A|\neg B)$ to measure the causal strength of B on A. This is intuitive for statistics, but more complex for psychology, requiring the introduction of **counterfactual reasoning** which we will discuss later. However, causal strength involves more than just ΔP. For example, if A is the music box starting to play, B is "an alien detects the object's position and makes the music box play", we will not believe it is plausible. The main reason of this is the hypothesis is too complex. In other words, we want the simplest hypothesis that explains the world. The belief is stronger when we use fewer variables to explain the effect. This is known as **Occam's razor**.

We also believe a relation is causal when it does not violate common sense. For example, we do not believe the Northeast blackout of 1965 was caused by a single handclap, even if that was the last event preceding the blackout.

2.3 Counterfactual reasoning

Perceiving causality is closely related to counterfactual reasoning. For example, if we regret not buying a lottery ticket, we imagine what would have happened if we had bought it. This counterfactual reasoning is believed to be a key distinction between causality and correlation. We cannot directly observe counterfactual outcomes for individuals, but can estimate statistically across groups. In the Rubin Causal Model (RCM) [9], we denote X_i as a binary variable for whether an individual is treated, $Y_i(x)$ as the outcome when $X_i = x$, and we can calculate the average causal effect (ACE):

$$\begin{aligned}\text{ACE}(X \to Y) &= E[Y_i(1) - Y_i(0)] \\ &= E[Y_i(1) - Y_i(0)|X = 1]P(X = 1) + E[Y_i(1) - Y_i(0)|X = 0]P(X = 0)\end{aligned}$$

Where $E[Y_i(1) - Y_i(0)|X = 1] = E[Y_i(1)|X = 1] - E[Y_i(0)|X = 1]$ is the average treatment effect which means the treatment effect on the treated. $E[Y_i = 0|X = 1]$ represents what would happen to originally treated people if they were not treated.

It is also important to note that counterfactual reasoning is not always accurate for human. For example, see Fig. 1. This further illustrates that causality is a highly subjective belief.

2.4 Causal factor and causal attention

As discussed above, humans favor the simplest hypothesis that explains observations. This aids our ability to discern causal factors among variables. For example, when a baby sees a music box play, the baby identifies precise causal factors (object position) while ignoring other variables (facial expression, sounds, etc). This ability is critical in constructing causal models given the complexity of real-world states and multitude of variables. As defined in [10], a causal factor is a variable that produces disjoint observation clusters when intervened on under a particular action sequence.

How can humans discern causal factors? First, we favor simple hypotheses, greatly narrowing the space. Second, we conduct controlled experiments to test hypotheses, as proposed in [10]. We also attend to what has changed and differs from the ordinary (linking to counterfactual reasoning)

Attending to what is novel and different is human nature, which allows us to rapidly build causal beliefs. Uncertainty arises from novel scenes and violation of expectation (VOE) events, motivating us to resolve the uncertainty. Cognitive studies [8] suggest an "adaptation" phenomenon where observers adapt to stimuli after prolonged viewing. This implies we preprocess visual data until eliminating uncertainty and perceiving causality.

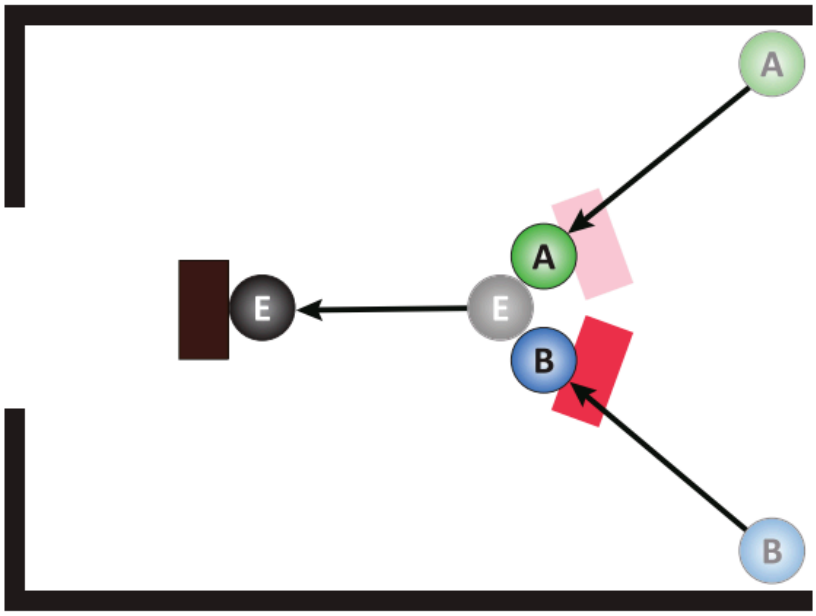

Figure 1: There are balls A, B, and E. A and B have some probability of going through the block. If and only if one ball goes through the block, ball E will go into the gate. The fact is that both balls go through the block.When asked whether A or B causes E to not enter the gate, most people will say B.However, the correct answer is that they contribute equally.Image credit: Gerstenberg and Icard [4].

3 Learning causal model in a new environment

While mathematical models of causality like SCM (Structural Causal Model) [7] and RCM (Rubin Causal Model) [9] are well-studied, learning causal models in new environments in a human-like way remains an open problem. Benchmarks such as CausalWorld [1] and OpenLock [2] have been proposed for causal structure learning. However, to enable human-like learning, we need open-ended, real world-like environments such as MINEDOJO (Fig. 2) by Fan et al. [3], where learning causality is not explicitly required but highly beneficial for multifaceted tasks. MINEDOJO is well-suited to simulate human causal learning given the diversity of hidden causal factors and complexity of the causal model. While landscapes and biomes vary, providing diverse tasks, the underlying causal structure remains stable. Moreover, low-level actions can leverage APIs like Mineflayer[1], enabling focus on high-level causal reasoning. Aside from RL, current MINEDOJO agents like Ghost [13] and Voyager [11] use large language models (LLMs) for planning with APIs for actions. Thus, an intuitive approach is to *build a causal model, guide LLM-based planning and acting, and learn via environmental exploration*.

As discussed above, the following features should be considered when building a causal model and enabling human-like learning:

- **Highly subjective**: Causal model building should be a highly subjective process. Humans do not need repeated experiences to establish causal relationships. We maintain a balance between changing hypotheses and conducting experiments to test them.
- **Strength of belief**: Causal strength represents confidence in the causal relationship. It should be related to the complexity of the hypothesis, common sense, and counterfactual reasoning. The strength should be updated through exploration.
- **Causal factors:** Discerning causal factors is essential given the complexity of the environment.
- **Causal attention:** To precisely locate causal relationships, the agent must be able to attend to what is special. It should also be equipped with memory and reflection abilities.

[1]https://github.com/PrismarineJS/mineflayer

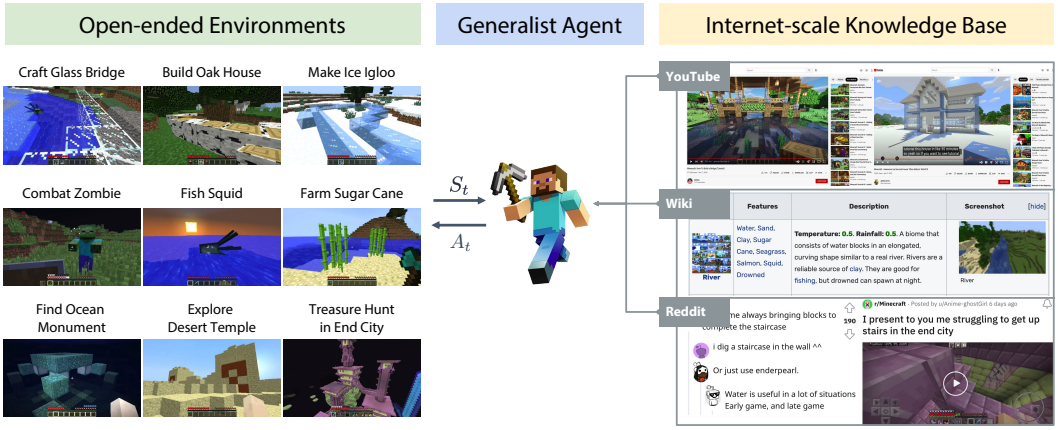

Figure 2: MINEDOJO a novel framework for developing open-ended, generally capable agents that can learn and adapt continually to new goals. Credit: Fan et al. [3].

4 Conclusion

In this essay, we discussed features of humans' causal learning abilities and suggested that when building an agent with such abilities, we should take these features into account. Being able to perceive causality is highly beneficial for building agents, as it helps solve problems in RL such as long-horizon planning, sparse rewards, *etc*. For example, in MINEDOJO, understanding *Hitting a block causes it to break* is essential for the agent to survive, but this is difficult without reward shaping. A causal model provides the agent with the ability to understand what it needs to complete a goal, which is a type of intrinsic motivation. This essay also suggests a brief pipeline for building an agent with causal learning abilities in MINEDOJO. Once a proper method to build a causal model is determined, this pipeline could construct an open-ended agent with life-long learning abilities that does not require complex reward shaping or prompt engineering.

References

[1] Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning, 2020. 3

[2] Mark Edmonds, James Kubricht, Colin Summers, Yixin Zhu, Brandon Rothrock, Song-Chun Zhu, and Hongjing Lu. Human causal transfer: Challenges for deep reinforcement learning. In *40th Annual Meeting of the Cognitive Science Society*, 2018. 3

[3] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL `https://openreview.net/forum?id=rc8o_j8I8PX`. 3, 4

[4] Tobias Gerstenberg and Thomas Icard. Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, 149(3):599, 2020. 3

[5] Alison Gopnik, David M Sobel, Laura E Schulz, and Clark Glymour. Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental psychology*, 37(5):620, 2001. 2

[6] Thomas L Griffiths and Joshua B Tenenbaum. Structure and strength in causal induction. *Cognitive psychology*, 51(4):334–384, 2005. 2

[7] Judea Pearl. *Causality*. Cambridge university press, 2009. 3

[8] Martin Rolfs, Michael Dambacher, and Patrick Cavanagh. Visual adaptation of the perception of causality. *Current Biology*, 23(3):250–254, 2013. 2

[9] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974. 2, 3

[10] Sumedh A Sontakke, Arash Mehrjou, Laurent Itti, and Bernhard Schölkopf. Causal curiosity: Rl agents discovering self-supervised experiments for causal representation learning. *arXiv preprint arXiv:2010.03110*, 2020. 2

[11] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. 3

[12] Wikipedia contributors. Humean definition of causality — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Humean_definition_of_causality&oldid=1154079866`, 2023. [Online; accessed 14-October-2023]. 1

[13] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory, 2023. 3