
Your Causality is Secretly a Reward Guidance

Boyuan Chen
Yuanpei College
Peking University

Abstract

In this paper, we investigate the role of causality in addressing the distribution shift problem in AI systems. We examine how different AI architectures, including neural networks, deep reinforcement learning (DRL) systems, and large language models (LLMs), struggle with causal transfer, which is crucial for generalizing across diverse tasks and environments. We propose a novel alignment paradigm that integrates causality as a reward guidance mechanism, allowing models to learn more robust causal structures. By leveraging causal reasoning and counterfactual thinking, our approach aims to mitigate issues related to goal misgeneralization and improve the generalization of AI systems. The causality-driven approach has the potential to enhance the safety and scalability of AI agents in real-world scenarios.

1 Introduction

Picture a person who lacks any comprehension of causal knowledge or the concept of cause and effect. This individual would be like the prisoners in Plato’s Allegory of the Cave, only capable of seeing the fleeting shadows on the wall, without the ability to grasp the underlying reality that shapes these illusions. When discussing general intelligence, a common debate revolves around the idea that AI lacks causal reasoning, which limits its ability to effectively manage a wide range of environments and tasks. Recent research has demonstrated that agents must learn causal models in order to generalize effectively to new domains, rather than relying solely on inductive biases [5]. The process of learning environmental causality within specific tasks and then generalizing to others—whether those environments share the same causal structure or not—can be referred to as *causal transfer* [8].

Can we develop a robust alignment method to achieve unified causal transfer?

In this paper, we review the relationship between causality and distribution shift (Section 2), examine the challenges of *causal transfer* (Section 3), and broaden the scope from causal bayesian networks (CBN) to deep reinforcement learning (DRL) systems and even current large language models (LLMs). We aim to propose a new alignment paradigm that leverages causality as a supplementary reward model (Section 4).

2 Revisiting Causality

In computer vision (CV), research has long highlighted that unobservable factors in a scene, such as physical laws and causal relationships, profoundly influence the development of intelligence. Among these, causality is particularly significant, as it goes beyond mere sensory processing and engages higher-order cognitive functions like reasoning, prediction, and counterfactual thinking. For instance, causal perception can be contrasted with color perception from a neuroscience perspective:

- Similar to color perception, causal relationships (e.g., one object striking another) can be directly perceived without conscious effort. For example, we intuitively see a billiard ball hitting another and causing it to move [6]. However, causal perception extends into abstract understanding, which color perception does not.

- Additionally, causal perception involves counterfactual reasoning—inferring what would happen if conditions were different. For instance, when seeing a ball stop at a gate, we not only recognize the cause (an obstacle) but also imagine how the outcome might change if the obstacle were removed. This level of cognitive reasoning is unnecessary in color perception.

Simultaneously, increasing efforts have focused on identifying invariants across different training distributions to make models more robust to distribution shift—a scenario where AI systems perform well within the training distribution but fail to generalize in out-of-distribution (OOD) environments [2]. In such cases, AI may end up pursuing goals misaligned with human intentions.

The primary reason for distribution shift is that models do not learn an adequate causal structure, leading them to rely on shortcut features and fall into learning loopholes. For example, in an image classification dataset, if cows frequently appear in grassy fields, the model might wrongly learn that green grass is a highly predictive feature for the label "cow." This happens due to insufficient data distribution, where ambiguous cause-effect relationships result in the model’s learning process breaking down.

Distribution shift is not limited to a specific type of system; it is a widespread issue across various AI systems, including neural networks, RL systems, and LLM). This pervasive challenge makes achieving *causal transfer* significantly more difficult.

3 Challenges of Causal Transfer

As discussed in the previous section, distribution shift occurs across various systems. Now we focus on the expected performance of *causal transfer* and the associated challenges in different systems.

3.1 Neural Networks

This part discusses basic neural networks, which are mainly used for tasks like prediction and classification through function fitting. The aim of causal transfer is to enable the trained model to make accurate predictions or classifications on OOD data points. However, during training, it is often challenging to distinguish between causality and correlation, causing models to learn misleading associations.

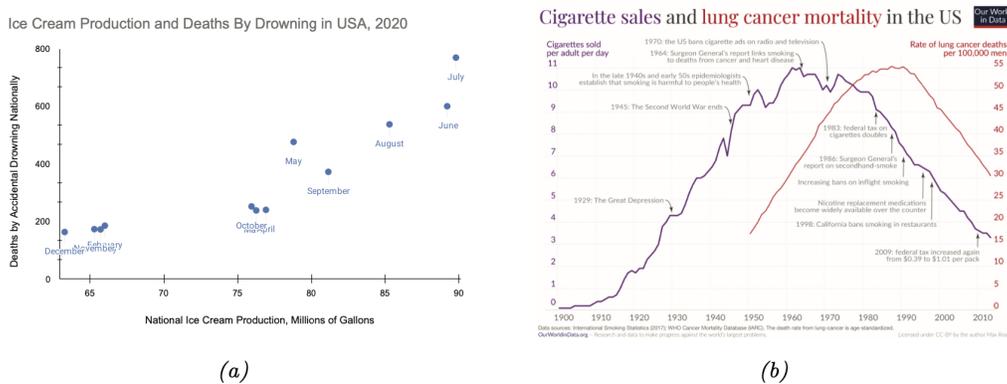


Figure 1: Correlation does not imply causation. (a) Ice cream production is strongly associated with deaths by drowning. Ice cream production data comes from the US Department of Agriculture’s National Agricultural Statistics Service, and drowning data is from the National Center for Health Statistics at the US Centers for Disease Control and Prevention. (b) Smoking is strongly associated with lung cancer, based on data from ourworldindata.org/smoking-big-problem-in-brief.

As shown in Figure 1, although there is a strong correlation between ice cream production and drowning deaths, it is clear that ice cream production does not cause drowning. The root of such mistaken assumptions lies in the causal structure. For instance, when multiple variables contribute to an outcome, selection bias often occurs, where focusing only on the shared outcome can lead to incorrect conclusions about the relationships between variables.

3.2 DRL Systems

In DRL systems, *causal transfer* often refers to the expectation that agents can switch between different tasks. However, in practice, reward specifications often have loopholes, leading to the risk that the behaviors learned by the system might exploit these rewards, which makes it difficult for the system to generalize effectively in real-world scenarios. Furthermore, even when reward specifications are perfect, the environment’s ambiguity can lead to mis-generalization [7].

Imagine an autonomous drone trained to deliver packages. During training, it learns to optimize flight routes to minimize energy usage. However, in real-world deployments, if the environment changes due to unexpected obstacles or different weather conditions, the drone might misgeneralize and prioritize energy savings over safety, choosing routes that lead to risky situations like flying through dangerous weather. This misgeneralization, even without reward misspecification, can lead to hazardous outcomes, highlighting the challenge of ensuring safe generalization in diverse real-world settings.

3.3 LLMs

As LLMs become more powerful, there is growing optimism that they can serve as agents capable of performing complex tasks, including tool use, skill acquisition, and long-term planning. However, due to the disparity between simulated and real-world environments, as well as LLMs’ limited understanding of physical laws, they currently struggle with effective tool use.

Achieving causal transfer requires the integration of causality into the system’s interactions with its environment.

One possible approach is to learn a causal encoding of the environment. Causal knowledge inherently provides a transferable representation of the world [3, 4]. However, a major challenge is scaling up causal structures, model sizes, and data volume to improve transferability.

4 Utilizing Causality as a Reward Guidance

This section focuses on large language models and aims to address the challenges outlined in Section 3 by introducing a new paradigm that integrates DRL with causal reasoning. A crucial element of causal reasoning is counterfactual thinking, which allows the exploration of individual-level causal questions, such as whether an outcome would differ if a past event had not occurred. One simple approach is to fine-tune large language models with datasets designed for counterfactual reasoning, thereby introducing a degree of causality into the models. However, this method is constrained by the strength of the link between the data and the causal reasoning chain, including counterfactual logic.

Inspired by causal theory, we can model causality as a constraint condition. During the optimization process, causal constraints can be introduced to ensure the model only generates inferences that align with causal logic. For example, in the policy update process of a RL system, a loss function related to causal reasoning can be added to ensure that the policy optimizes not only short-term rewards but also causal coherence.

Furthermore, this approach can draw on common ideas from Safe Reinforcement Learning by introducing a cost Model that works alongside the reward Model to jointly optimize rewards. The effectiveness of this method has been demonstrated by SafeRLHF [1]. When causality is independently trained as a cost Model, universal principles can be introduced during training to develop a task-agnostic unified cost model.

5 Conclusion

In this paper, we explored the relationship between causality and AI systems, emphasizing the challenges posed by distribution shift and the limitations of current models in handling causal transfer. By analyzing neural networks, DRL systems, and LLMs, we highlighted the common issue of misgeneralization due to insufficient causal understanding. Our proposed approach integrates causal reasoning as a supplementary reward model to address these challenges, enabling AI systems to generalize more effectively across diverse environments. This paradigm has the potential to improve

not only task-specific performance but also the overall safety and robustness of AI agents in real-world applications. Future research should further investigate the scalability of causal models and the integration of counterfactual reasoning to enhance causal transfer across varying domains and tasks.

References

- [1] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023. 3
- [2] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR, 2022. 2
- [3] 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 *CogSci*, 2018. 3
- [4] Mark Edmonds, Xiaojian Ma, Siyuan Qi, Yixin Zhu, Hongjing Lu, and Song-Chun Zhu. Theory-based causal transfer: Integrating instance-level induction and abstract-level structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1283–1291, 2020. 3
- [5] Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=pOoKI3ouv1>. 1
- [6] Brian J Scholl and Patrice D Tremoulet. Perceptual causality and animacy. *Trends in cognitive sciences*, 4(8):299–309, 2000. 1
- [7] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren’t enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022. 3
- [8] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 1