Can Computers Think?
A Causal Perspective

Tobias Maringgele

Introduction

The question of whether computers can "think" has long captured attention in computer science and philosophy. Alan Turing proposed a testable criterion for machine intelligence, known as the Turing Test, which focuses on whether machine behavior is indistinguishable from a human's [Turing, 1950]. John Searle later countered this view with his Chinese Room argument, which questioned whether symbolic manipulation alone constitutes understanding [Searle, 1980]. While these foundational discussions emphasize behavioral or linguistic mimicry, this essay explores a more structural dimension: whether machines can reason about cause and effect. Drawing on Judea Pearl's theory of causal inference [Pearl, 2009, Pearl and Mackenzie, 2018], I argue that thinking—at least in the human sense—depends crucially on the ability to engage in counterfactual reasoning. Most AI systems today operate purely on statistical correlation; they fall short of what we typically mean by "thinking," as they cannot reason about interventions or imagine alternative outcomes. This essay analyzes the limitations of current AI and how causal models offer a path toward more meaningful machine cognition.

Limitations of Correlational AI

Most of today's AI systems, particularly those using deep learning, operate on vast amounts of data to discover patterns of correlation, but not the underlying mechanisms that generate them. They excel at tasks such as image recognition, speech transcription, and language modeling. However, these systems typically lack an understanding of why certain patterns exist. As

Pearl and Mackenzie describe, these models are limited to the "first rung" of the *Ladder of Causation*: pure association [Pearl and Mackenzie, 2018]. They can answer questions like, "What is the probability of Y given X?", but not "What would have happened if we had done Z instead?" This limits their generalization capabilities, particularly when facing scenarios that deviate from the training distribution. Without a model of the underlying causal structure, AI systems may perform well on benchmark datasets yet fail dramatically when deployed in real-world environments where interventions and confounding factors abound.

The Ladder of Causation

Pearl's Ladder of Causation illustrates the increasing cognitive demands of different kinds of questions—moving from mere observation to reasoning about actions and imagined alternatives. His framework distinguishes between three levels of reasoning—with only the upper two involving causal inference: association, intervention, and counterfactuals [Pearl, 2009, Pearl and Mackenzie, 2018]. Most machine learning systems operate at the associative level, inferring patterns from observed data. Intervention, the second level, involves reasoning about the effects of actions—answering questions like, "What will happen if we do X?" This is essential in fields such as robotics or medicine, where agents interact with the world and must anticipate the consequences of their interventions. The third and highest level involves counterfactual reasoning: considering alternate scenarios such as, "Would the outcome have been different if I had done A instead of B?" Each level builds on the previous, with counterfactual reasoning requiring a grasp of both interventions and associations.

Humans naturally operate on all three levels. We recognize patterns, perform actions, and evaluate alternate realities. This ability to simulate and learn from unobserved scenarios is arguably a hallmark of human intelligence. Current AI systems largely lack access to these higher levels of the ladder, and thus miss out on core components of what it means to "think" in the human sense.

Counterfactuals and Thinking

The ability to reason counterfactually is fundamental to planning, ethical reasoning, and scientific inquiry. Cognitive science research shows that humans create alternatives to reality through structured, rational processes that support moral judgment, learning, and decision-making [Byrne, 2007]. We frequently engage in hypothetical thinking: "If I had studied more, I would have passed the exam," or "If I wrote a better essay, I would have been accepted to the program." These are not just speculative thoughts; they guide decisions, assign responsibility, and facilitate learning from non-actualized events.

Causal models, especially *structural causal models* (SCMs), enable formal reasoning over counterfactuals by explicitly representing the underlying mechanisms that generate data. These models allow an agent to answer "what-if" questions by modifying structural equations and computing the resulting outcomes [Pearl, 2009]. Recent advances have also made it increasingly feasible to learn such models from data. For instance, Pawlowski et al. propose a deep generative framework that supports tractable counterfactual inference by combining SCMs with normalizing flows and variational inference [Pawlowski et al., 2020]. This line of work illustrates that counterfactual reasoning can be implemented computationally, offering a path toward more flexible and interpretable AI systems.

Conclusion

The question "Can computers think?" hinges on how we define thinking. If we mean the ability to mimic human output, current systems offer partial success. But if we define thinking as the capacity for flexible, hypothetical, and causal reasoning, then today's machines fall short. Causal inference—especially counterfactual reasoning—is not an optional add-on but a core component of cognition. Pearl's causal ladder provides both a theoretical foundation and a practical roadmap for building systems that more closely mirror human thought. Moreover, combining data-driven learning with causal inference promises AI systems that not only recognize patterns but also understand and explain the mechanisms behind them.

In embracing causal modeling, we take steps toward machines that do not

just see, but understand; not just act, but explain; not just learn, but imagine. While consciousness and subjective experience remain out of reach—and are arguably not necessary for engineering intelligent systems—it is through causality that computers may begin to think in a meaningful, useful sense. As research progresses, causal reasoning is likely to become a foundational element in the development of more robust, generalizable, and transparent AI systems.

References

Ruth M. J. Byrne. Précis of the rational imagination: How people create alternatives to reality. *Behavioral and Brain Sciences*, 30(5-6):439–453, 2007. doi: 10.1017/S0140525X07002586.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980.

Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236): 433–460, 1950.