

MIT Press • Open Encyclopedia of Cognitive Science

Causal Reasoning

David Lagnado¹

¹Dept. of Experimental Psychology University College London

MIT Press

Published on: Sep 30, 2024

DOI: <https://doi.org/10.21428/e2759450.0de13fd2>

License: [Creative Commons Attribution-NonCommercial 4.0 International License \(CC-BY-NC 4.0\)](#).

We engage in causal reasoning on a daily basis, whether we are stacking books, figuring out why a friend is upset, or deciding the guilt of a defendant. It allows us to predict and explain the world around us, to imagine what-if, and to assign praise and blame. What differentiates causal reasoning from other forms of inference is that it not only concerns what actually happens but also requires us to imagine how things might have been different. To establish if X causes Y, we need to infer whether Y would still have occurred without X. How do people achieve this imaginative feat? An emerging view in cognitive science is that people build internal causal models that allow them to simulate possible scenarios and hence to compare what actually happened against other counterfactual possibilities. For example, to decide whether eating the seafood caused my rash, I mentally simulate alternative scenarios where I didn't eat the seafood and see if the rash still occurs. This general-purpose capability allows us to reason about both physical and social systems and also underpins our ability to construct stories and explanations.

History

While the nature of causality is still debated among philosophers, all agree that it is fundamental to how people construe the world. But how do people engage in cause-and-effect reasoning? A central claim is that we use mental models to represent and reason about the world. The key idea is that by manipulating our mental models, we can predict how things will change. Early precursors were cognitive maps that represent spatial structure ([Tolman & Brunswik, 1935](#)) and internal models that “run in the mind” to predict the consequences of our actions ([Craik, 1952](#)). The next iteration of mental models was logic-based, primarily designed to explain deductive reasoning ([Johnson-Laird, 1983](#)). However, the logical frame underpinning this approach struggles to capture crucial elements of causal reasoning ([Sloman & Lagnado, 2005](#)).

A critical breakthrough for cognitive science was the advent of causal Bayes nets ([Pearl, 2000](#)), which formalize the concept of a causal model and provide a computational framework for causal and counterfactual inference [see [Causal Learning](#)]. It distinguishes inferences based on observation (“I see the lights are on and infer that someone is home”) from inferences based on interventions (“I switch the lights on but do not infer that someone else is home”) and defines counterfactuals in terms of interventions on causal models (see [Figure 1](#)). It also links causal models with probabilities and shows how Bayesian updating depends on causal structure.

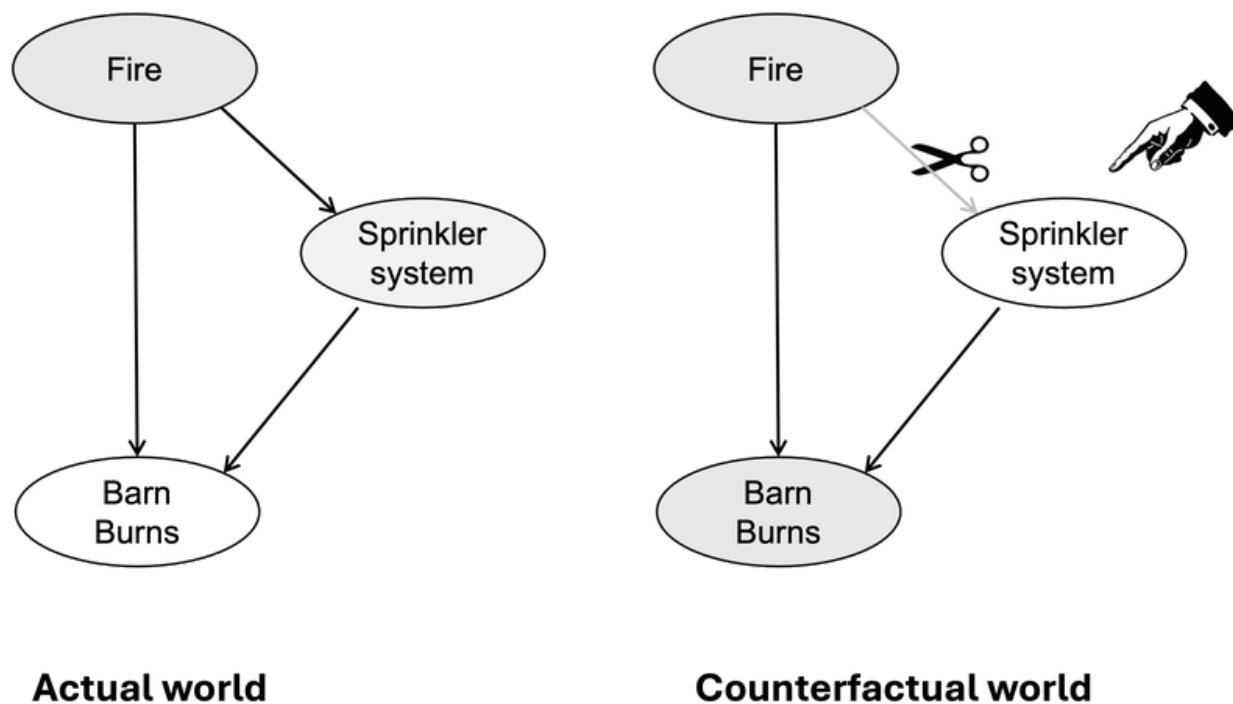


Figure 1
Using a causal model for counterfactual inference.

Suppose that a small fire starts in a barn, but the sprinkler system activates, preventing the barn from burning. Would the barn have burned down if the sprinkler system did not work? This question requires a reasoner to imagine a counterfactual world where the sprinkler does not activate and to then predict the probability of the barn burning. But the reasoner should not infer that because the sprinkler did not activate the fire did not start (which would only be legitimate if they had merely observed that the sprinkler did not activate). The causal model framework shows the correct inference in such cases, and human reasoners follow this pattern of reasoning ([Sloman & Lagnado, 2005, 2015](#)).

This framework sparked a wealth of novel research on causal reasoning, serving as a benchmark against which to compare human judgment but also as a guide to the kinds of representations and inferences that people might use ([Sloman, 2009](#)). New questions were addressed, such as how people use causal models for probability judgments ([Krynski & Tenenbaum, 2007](#)), how they adapt their models for counterfactual inference ([Lucas & Kemp, 2015](#); [Sloman & Lagnado, 2005](#)), how they use causal knowledge to categorize new objects ([Rehder, 2003](#)), and how they generate causal models from prior knowledge ([Griffiths & Tenenbaum, 2009](#)).

A key question is the nature of the cognitive apparatus that allows people to construct causal models and run mental simulations. One proposal is that when people reason about everyday physical situations—such as predicting whether a pile of plates will topple—they use their intuitive knowledge of physics to simulate possible trajectories and interactions of the objects ([Battaglia et al., 2013](#); [Ullman et al., 2017](#)). A

naïve physics engine in the head enables a range of causal and counterfactual inferences, allowing people to judge what would happen if the system was perturbed in various ways ([Gerstenberg et al., 2021](#)). Further evidence that people simulate counterfactuals when making causal judgments is given by eye-tracking ([Gerstenberg et al., 2017](#)): When judging causality in object collisions, participants not only track the actual motion of the objects but also track where they would have gone but for the collision.

Across a wide range of tasks, people show an impressive capacity for causal reasoning. While some psychological theories cast causal simulation as a short-cut heuristic or bias ([Kahneman & Tversky, 1982](#)), the new approach argues that causal thinking is a foundational part of reasoning, not a rule of thumb. This does not mean that people are immune to error and biases. But heuristics enter when people need to use approximate inference because of the complexity of the problem and their limited computational resources ([Bramley et al., 2017](#)).

Core concepts

Causal mental models are internal representations of the world that capture causal structure and allow people to predict and explain how things change. What distinguishes these models from associative or logic-based models is that they can be used to predict the consequences of novel interventions. For example, my causal model of my cat allows me to predict what will happen if I tread on its tail.

Counterfactuals are imagined alternatives to what actually happened. For example, suppose you take the bus to work and arrive late. What would have happened if you had cycled instead? Counterfactuals can be tricky because we need to decide what to hold constant from the actual world (e.g., the weather conditions) and what to allow to change (e.g., the route taken). The ability to construct plausible counterfactuals is a strength of human reasoning, but the details of how we achieve this are still under debate.

Mental simulation is the running of an internal model, from initial conditions through to possible outcomes, that supports prediction and can also be used to evaluate counterfactuals. For example, to predict what would happen if I cycle rather than take the bus, I mentally simulate cycling to work and assess how it is affected by the traffic conditions.

Intuitive theories are commonsense knowledge of a domain, such as the physics of everyday objects or social interactions, that allows us to generate plausible causal models for specific situations [see [Intuitive Theories](#)]. For example, my intuitive theory about pets allows me to generate a causal model of the behavior of my cat, which in turn supports prediction via simulation.

Questions, controversies, and new developments

One challenge for the mental simulation account is how to scale up to complex situations or those involving abstract concepts. For example, how might one use mental simulation to address causal questions about the economy or the efficacy of military interventions? A related problem is that people

show distinctive patterns of errors in physical reasoning that seem incompatible with the use of mental simulation ([Ludwin-Peery et al., 2021](#)). A lot depends on how simulation is characterized and how much it relies on high-fidelity versus good-enough processing. One suggestion is that people use partial rather than full simulation because of limited computational resources ([Bass et al., 2021](#)).

Another challenge is how people decide which counterfactuals to use and how to avoid problems such as over-determination (where several causes are each sufficient for an effect) or pre-emption (where one cause pre-empts an alternative cause). In both cases, a simple counterfactual will deliver the wrong causal judgment ([Halpern, 2016](#)). Various attempts have been made to circumvent these problems, but so far, no agreed-upon cognitive or formal account captures all the nuances.

One exciting future direction is exploring how people create novel causal concepts by combining and reusing previously acquired knowledge ([Zhao et al., 2023](#)). The idea that despite our cognitive limitations we can use bootstrapping to invent richer causal concepts and also share this knowledge with others, opens the door for a deeper understanding of how we develop and transmit causal knowledge.

Broader connections

Causal reasoning lies at the heart of many practical areas of decision making, such as law, health, and the environment. By better understanding how people use causal models in these complex domains, the quality of people's decision making may be improved. For example, the dominant psychology model of legal decision-making is predicated on the claim that people use causal knowledge to build stories that explain evidence and determine verdicts ([Lagnado, 2021](#); [Pennington & Hastie, 1992](#)). But exactly how people achieve this, and how well it corresponds to normative causal inference, remains under-studied.

Causal reasoning is also a precursor to assigning responsibility, whether moral or legal. Inspired by the causal simulation approach, research is starting to unpack the factors that determine how people attribute responsibility when multiple agents contribute to critical outcomes, such as the success or failure of a joint project ([Lagnado et al., 2013](#)). This work extends to understanding how to encourage people to take more responsibility for their actions.

Further reading

- Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. MIT Press. <https://doi.org/10.7551/mitpress/9540.001.0001>
- Sloman, S. A. (2009). *Causal models: How people think about the world and its alternatives*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195183115.001.0001>
- Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, 66(1), 223–247. <https://doi.org/10.1146/annurev-psych-010814-015135>
- Waldmann, M. R. (Ed.). (2017). *The Oxford handbook of causal reasoning*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199399550.001.0001>

References

- Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, 38(7-8), 413–424.

<https://doi.org/10.1080/02643294.2022.2083950>

↩

- Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 18327–18332.

<https://doi.org/10.1073/pnas.1306572110>

↩

- Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, 124(3), 301–338.

<https://doi.org/10.1037/rev0000061>

↩

- Craik, K. J. W. (1952). *The nature of explanation*. Cambridge University Press.

↩

- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, 128(5), 936–975.

<https://doi.org/10.1037/rev0000281>

↩

- Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, 28(12), 1731–1744. <https://doi.org/10.1177/0956797617713053>

↩

- Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, 116(4), 661–716. <https://doi.org/10.1037/a0017201>

↩

- Halpern, J. Y. (2016). *Actual causality*. MIT Press. <https://doi.org/10.7551/mitpress/10809.001.0001>

↩

- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

↩

- Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge University Press. <https://doi.org/10.1017/CBO9780511809477.015>

↩

- Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3), 430–450. <https://doi.org/10.1037/0096-3445.136.3.430>

↩

- Lagnado, D. A. (2021). *Explaining the evidence: How the mind investigates the world*. Cambridge University Press. <https://doi.org/10.1017/9780511794520>

↩

- Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, 37(6), 1036–1073. <https://doi.org/10.1111/cogs.12054>

↩

- Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, 122(4), 700–734.

↩

- Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, 127, 101396. <https://doi.org/10.1016/j.cogpsych.2021.101396>

↩

- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

↩

- Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology*, 62(2), 189–206. <https://doi.org/10.1037/0022-3514.62.2.189>

↩

- Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, 27(5), 709–748. https://doi.org/10.1207/s15516709cog2705_2

↩

- Sloman, S. A. (2009). *Causal models: How people think about the world and its alternatives*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195183115.001.0001>

↩

- Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, 66(1), 223–247. <https://doi.org/10.1146/annurev-psych-010814-015135>

↩

- Sloman, S. A., & Lagnado, D. A. (2005). Do we “do”? *Cognitive Science*, 29(1), 5–39. https://doi.org/10.1207/s15516709cog2901_2

↩

- Tolman, E. C., & Brunswik, E. (1935). The organism and the causal texture of the environment. *Psychological Review*, 42(1), 43–77. <https://doi.org/10.1037/h0062156>

↩

- Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 21(9), 649–665. <https://doi.org/10.1016/j.tics.2017.05.012>

↩

- Zhao, B., Lucas, C. G., & Bramley, N. R. (2024). A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, 8(1), 125–136. <https://doi.org/10.1038/s41562-023-01719-1>

↩