Philosophy of Science (2026), 1–10
doi:10.1017/psa.2026.10285

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

SYMPOSIA PAPER

Empowerment Gain as Causal Learning, Causal Learning as Empowerment Gain: A Bridge between Bayesian Causal Hypothesis Testing and Reinforcement Learning

Alison Gopnik

Department of Psychology, University of California Berkeley, Berkeley, USA
Email: gopnik@berkeley.edu

(Received 29 December 2024 UTC; revised 19 August 2026 UTC; accepted 20 August 2026 UTC)

Abstract

Cognitive scientists have applied advances in our formal understanding of causation in philosophy and computer science, particularly within the Causal Bayes Net formalism, to understand human causal learning. In parallel, reinforcement learning (RL) researchers have developed the idea of an intrinsic reward signal called "empowerment." An agent is rewarded for maximizing the mutual information between its actions and their outcomes, regardless of the external reward value of those outcomes. Empowerment may be an important bridge between classical Bayesian causal learning and RL and may help to characterize causal learning in humans and enable it in machines.

Learning about the causal structure of the world is a fundamental problem for human cognition, and causal knowledge is central to both intuitive and scientific theories. Over the last twenty years or so, cognitive scientists have applied advances in our formal understanding of causation in philosophy and computer science to understand human causal learning. In particular, researchers have used the Causal Bayes Net formalism, which relates directed acyclic causal graphs to patterns of conditional probability, interventions, and counterfactuals in systematic ways (Pearl 2000 (2009); Spirtes et al. 2001). The formalism assumes an "interventionist" account of causation —roughly, variable X is causally related to variable Y if, holding all else constant, an intervention changing the value of X would lead to a change in the value of Y. It provides a natural way to describe both causal models and the patterns of data they generate. This allows the reverse Bayesian inference, at least in principle—from the pattern of data it should be possible to infer the causal model that was most likely to have generated that data. This approach to causal learning is part of a broader

© The Author(s), 2026. Published by Cambridge University Press on behalf of Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psa.2026.10285 Published online by Cambridge University Press

Bayesian approach to learning in developmental cognitive science (for a recent review see Ullman and Tenenbaum 2020). Empirically, causal learning in children is remarkably well captured by this formalism (for reviews see Gopnik and Wellman 2012; Gopnik and Bonawitz 2015; Goddu and Gopnik 2024). Given a pattern of data, even preschool children infer the correct causal hypothesis.

From a computational perspective, however, this kind of inference, like many other kinds of Bayesian inference, proves to be intractable—the space of possible hypotheses is too large. Various computational techniques have been used to try to deal with this problem, particularly sampling methods, and there is some evidence that young children use similar methods (Bonawitz et al. 2014a, 2014b). These methods have had some success, but the search problems remain challenging. Another approach might be to suggest that potential causal models are strongly constrained by prior innate "core knowledge" (e.g., Spelke 2022). Although this might help address the search problem, it also undermines one of the major advantages of causal learning. Causal learning is so valuable precisely because it allows us to go beyond innate knowledge and learn about causal structure that may be deeply counter intuitive—such as the causal structure of scientific theories.

Another challenge for Bayesian approaches involves the role of exploration, experimentation, and active learning. Children are not simply passive consumers of data. Instead from very early on they actively seek out evidence that is relevant to the causal problems they are trying to solve, and this exploration plays an important role in their solutions to those problems. There has recently been extensive and elegant work showing just how motivated children are to experiment and explore, and how intelligently they do so (e.g., Schulz 2012; Giron et al. 2023). Similarly, in formal science, experimentation is the canonical way to discover causal structure. But, with a few exceptions, (e.g., Eberhardt and Scheines 2007) there have not been computational accounts of how this kind of active intervention and experimentation takes place and how it allows causal inference. So, the cognitive science of causal learning seems to have reached a kind of impasse.

However, in parallel, an apparently quite different kind of learning mechanism—reinforcement learning (RL)—has become increasingly influential in neuroscience and computer science, particularly when combined with modern "deep" machine learning (Dayan and Balleine 2002; Sutton and Barto 2018). For example, deep RL has been the key to Deep Mind's accomplishments in mastering Go and Chess (Silver et al. 2018).

As the philosopher James Woodward, among others, has pointed out, classical RL might be thought of as a very specific and narrow form of causal learning (Woodward 2007, 2021). In particular, it relates the specific actions an agent performs on the world—their interventions—to specific outcomes, in the form of external rewards. This basic structure is similar to the basic structure of causal relations on the interventionist account underlying the causal Bayes net formalism. Moreover, RL contrasts with other types of learning that are more purely associative, such as classical conditioning or many neural network learning mechanisms. Associative learning captures correlations between variables and allows predictions about those variables, but doesn't allow a special role for interventions, and so is further removed from causal learning.

https://doi.org/10.1017/psa.2026.10285 Published online by Cambridge University Press

However, RL is much narrower in application than causal learning more generally, it typically applies only to the agent's own actions and to the rewards that follow those actions. More importantly, the basic motivational structure of RL and causal learning are very different. Reinforcement learning is motivated by utilities—the attempt to maximize external rewards. Classically, causal learning is epistemically motivated—it involves approximating the true structure of the world. It is true that ultimately causal knowledge allows effective action and decision making and so increases utilities. But this is a long-term and indirect effect. In the Bayesian framework, decision making and utility calculations are layered on top of the fundamental epistemic project (as in "causal decision theory"). The reverse is true in RL; RL agents may try to learn about the environment but they only do so in service of the fundamental utility project—that is, they learn to maximize further rewards later.

Although RL methods can be very effective in relatively low-dimensional and fixed environments such as go or chess, especially with dense reward signals, they have much more difficulty in the high-dimensional and open-ended environments that are characteristic of human causal learning. In particular, RL systems face consistent difficulties in balancing exploitation—the fundamental utilitarian motivation of maximizing reward, and exploration—the more epistemic motivation of learning about the structure of the environment (e.g., Sutton and Barto 2018; Cohen et al. 2007). In the long run, exploration will lead to more effective action and reward, but it requires the agent to forego reward in the short run. These problems arise even in very simple "bandit tasks" where agents are only required to choose between actions with known or unknown outcomes. Choosing the known option allows an agent to be sure of reward, but the unknown option will be more informative and lead to more knowledge and so to more effective action in the long run.

One approach to solving these problems has been to propose systems that seek internal epistemic rewards in addition to the classical external rewards. These systems implement the intrinsic epistemic motivation of classical causal learning in an RL framework. Several types of intrinsic rewards have been proposed in the literature. They include a variety of curiosity-based rewards, particularly measures of information gain, entropy, and novelty (Oudeyer et al. 2007; Schmidhuber 2010; Pathak et al. 2017). These rewards do seem to increase the exploratory efficacy of RL systems. Moreover, there is evidence that children and even infants also seek out such intrinsic rewards, particularly information gain (Schulz 2012; Kidd and Hayden 2015).

However, it is still difficult to sufficiently constrain these systems. For example, a classic failure mode for systems that seek out information gain is the "noisy TV" problem (Schmidhuber 2010). Such systems will be captured by random noise like TV static. Seeking novelty and information gain by themselves doesn't seem to be adequate to understand the environment effectively in a way that supports action.

A different approach uses "empowerment" as an intrinsic epistemic reward (Klyubin et al. 2005; Salge et al. 2014). "Empowerment" originally emerged in the evolutionary computation and artificial life literature, but more recently it has been applied to a variety of RL problems (Du et al. 2020; Zhao et al. 2021; De Abril and Kanai 2018). In empowerment, an agent maximizes the mutual information between its actions and their outcomes, regardless of the reward value of those outcomes. In other words, the system is rewarded if variation in an action systematically leads to parallel variation in the outcome so that the value of the action predicts the value of

https://doi.org/10.1017/psa.2026.10285 Published online by Cambridge University Press

the outcome. Simultaneously, the system is also rewarded for maximizing the variety of actions it takes, ensuring that it explores the widest range of possible actions and outcomes.

The result is an intrinsic motivation system that leads the agent to seek the maximum amount of control over its environment. By endowing RL agents with empowerment as an intrinsic reward those agents can explore and represent the environment more effectively. Rather than simply seeking out novelty or information, such a system will seek out exactly the relations in the world that involve the closest match between actions and outcomes—the most controllable relations. This means that it will discover the relations that will be most likely to be useful for a wide range of future goal-directed actions. If I discover, for example, that moving a stick systematically changes the position of objects that it contacts, I can later pick up the stick to draw an out of reach object toward me. In fact, infants seem to learn to use sticks in just this way (Uzgiris and Hunt 1975).

Although they come from different traditions, I argue that causal learning and empowerment gain are intimately related. In particular, my hypothesis is that if an agent learns an accurate causal model of the world, they will necessarily increase their empowerment, and, vice versa, increasing empowerment will lead to a more accurate (if implicit) causal model of the world. This has implications for both cognitive science accounts of causal learning and for the metaphysics of causation.

This claim is rooted in the peculiar ontology of causation. A number of researchers have suggested that it is helpful to think of causal learning as an "inverse problem" (e.g., Gopnik et al. 2004). An inverse problem involves inferring the structure of the external world from the data that world generates. A classic example is the way a visual system infers the structure of the three-dimensional world from retinal images or pixels. Other examples might include the way that we infer beliefs and desires from behavior in "theory of mind," infer neural structure from fMRI evidence in cognitive neuroscience, or infer atomic structure from cloud chamber traces. In all these cases we assume that there is, in fact, some single distinctive structure in the outside world that we are trying to reconstruct. A God's-eye view of the universe would discover something corresponding to the 3-D structure or psychological structure or neural structure or atomic structure that we are trying to reconstruct. This structure is independent of the agents who are trying to understand it.

However, the most influential contemporary accounts of causation suggest a rather different ontological picture of causal relationships. The interventionist accounts that underlie causal Bayes nets (Woodward 2003; Pearl 2000 (2009); Pearl and Mackenzie 2018; Spirtes et al. 2001) imply a different relation between agents and the world. The asymmetries between cause and effect that are central to interventionist accounts of causation are, notoriously, hard to be found in physics, at least at the micro level. Instead, our notions of causation are rooted in the idea that causal relations are precisely those external relations that support an agent's interventions. Philosophers and computationalists in this tradition define causal relations as those relations such that intervening to alter the value of a cause variable will lead to a corresponding change in the effect variable—in short, those relations where actions predictably produce outcomes. The ideal interventions that underpin causal inference are not identical to the intentional actions of actual agents but they are closely related, and in many circumstances agents' intentional actions will also serve as ideal

https://doi.org/10.1017/psa.2026.10285 Published online by Cambridge University Press

interventions (see Woodward 2021) and so pick out causal relationships in the world. Discovering such relations between interventions and outcomes is also the fundamental idea behind empowerment. But there may not be any single God's-eye view agent-independent causal structure analogous to, say, 3-D spatial structure or atomic structure.

The philosopher Peter Godfrey-Smith (2010), among others, has argued for "causal pluralism"—in which many ontologically disparate phenomena can all support causal interventions. On this view there is nothing analogous to the spatial structure of the 3-D world in the causal case. Rather there are many quite different relationships in the world, ranging from the intuitive physical relations that support "billiard-ball" causation to the belief-desire relationships of intuitive psychology, to the highly counterintuitive relations of relativistic physics, that all happen to systematically support causal interventions. Again, these are precisely the relations of control that would produce the greatest increases in empowerment. Maximizing empowerment will lead to the discovery of causal relations, and vice-versa.

A further distinctive feature of empowerment, with implications for the metaphysics of causation, is that it can have both "mind to world" and "world to mind" directions of fit. It is possible to maximize an agent's empowerment by increasing its knowledge about how the world works (matching the mind to the world as in science). This is the classic picture of causal learning. However, it is equally possible to maximize empowerment by increasing an agent's skill and control over the world (matching the world to the mind as in engineering). This is more like the orientation of classic RL.

These engineering cases point to other distinctive features of the metaphysics of causation. In these cases, we arguably create causation as well as discovering it. Human artifacts are a good example of this "world to mind" causal fit. Designing a gas pedal imposes a relation of mutual information between the pedal and the car acceleration that did not exist before the car was engineered. Arguably, also, it creates a causal relationship between the pedal and the acceleration that did not exist before. This is even true for simpler forms of tool use like using a stick to obtain an out of reach object. The stick lying on the forest floor might not have any causal relation to toy retrieval, but the stick as a tool in the hands of a toddler does.

Something similar happens in the relation between norms and "theory of mind." By imposing a new social norm or convention we can introduce novel causal relations to the psychological as well as the physical world. By sending a single email that class starts at 4:00 in room 100 I can miraculously and precisely cause a hundred separate people to converge at just that place and time.

However, whether we think in terms of science or engineering, the basic structure of causal relations will be the same, intervene systematically on X to systematically change the value of Y. This is well captured by the notion of empowerment. Of course, in more abstract and conceptual cases, the interventions may be theoretical rather than actual—to say that the moon causes the tides is to say that if we could alter the position of the moon the tides would also alter, even if this intervention isn't possible. But, as Woodward argues (2003, 2021), if a relation is genuinely causal this sort of intervention should at least be conceptually possible. Moreover, this distinguishes causal relationships from other relationships such as spatial, geometric, or logical relationships. And, in practice, the test for whether we have discovered causal

https://doi.org/10.1017/psa.2026.10285 Published online by Cambridge University Press

relations is to perform experiments—to determine whether experimentally varying one variable will systematically predict the value of another, that is, precisely to look for high mutual information between interventions and outcomes.

Thinking about empowerment might also help us understand the psychology of causal learning. A recent paper suggests that adult's exploration of a video game-like environment can be well characterized by empowerment (Brändle et al. 2023) and we have shown that this is also true for children (Du et al. 2023; Yiu et al. 2026). But the empowerment approach more generally captures important features of early causal knowledge and learning and helps to explain a wide range of developmental findings.

Looking-time studies suggest that very young infants perceive some particular relations in intuitive physics that support causal inference, such as the relations of movement and collision in "billiard-ball" causality (Leslie 1982). However, the development of causal concepts more broadly is initially closely linked to actual goal-directed actions on the world and their outcomes. Goddu and Gopnik (2024) following Woodward (2007, 2021) suggest that both in phylogeny and ontology, causal understanding moves from a first-person perspective to a third-person perspective to an impersonal perspective. Reinforcement learning, which is found in almost all intelligent animals, represents a causal relation between the animal's own actions and their outcomes. Imitation learning, which is found in some nonhuman animals in limited ways, but is ubiquitous in human infants, represents a causal relation between other animals' actions and outcomes. The sort of impersonal causal understanding in science represents causal relations in the world independent of actual actions, though crucially supporting such actions in principal. Empowerment may be applied to all three types of relations.

In the 1970s, and interestingly in the context of thinking about operant conditioning, a series of papers suggested that even very young infants are indeed rewarded by something like empowerment. In classic studies of "conjugate reinforcement," Rovee-Collier (1979) tied a ribbon from a crib mobile to the infant's foot, so that kicking made the mobile move. Infants as young as three months old systematically acted to make the mobile move, varying their actions and observing the correlation between those actions and the behavior of the mobile. There were similar results in studies where infants could make a mobile move or activate a pattern of lights by turning their heads on a pressure sensitive pillow (Watson 1972; Papousek and Papousek 1975). Moreover, these actions could not simply be explained by classic RL with the mobile's motion as a reward. Infants would learn to turn their heads or kick and would continue to act to do so, even though they no longer looked at the mobile or the lights, aside from a brief glance to check that their action was effective. Infants varied their actions and observed their results rather than simply converging on a single effective action. In addition, infants smiled and cooed when their actions consistently led to an effect, but not when that effect simply occurred independently of their actions (Watson 1972). A more recent study with this methodology shows that infants consistently alter their actions on the mobile in a way that increases the contingency of their actions, again unlike classical RL (Sloan et al. 2023). In fact, Rovee-Collier described her results precisely as an empowerment reward: "The control which the infants have gained over the consequences of their own actions seems to have become the reward, rather than the specific consequences per se." (Rovee-Collier and Gekoski 1979).

https://doi.org/10.1017/psa.2026.10285 Published online by Cambridge University Press

Later in infancy infants even more clearly act on intermediate objects to maximize empowerment—looking for contingencies between potential tools and outcomes, and at least implicitly appreciating the causal relations between them. Piaget (1952) described "tertiary" circular reactions in infancy. In these behaviors, which emerge at around a year, infants systematically vary their actions on objects to produce varied outcomes. Again, Piaget explicitly described this behavior in terms of experimentation and suggested that it was a precursor to causal understanding, though for complicated reasons he denied that children had genuine causal understanding until well into the school-aged years.

In "conjugate reinforcement" infants are acting to maximize the empowerment of their own actions—their causal understanding has a first-person perspective. We know that from early in infancy children also represent the goal-directed actions of others and distinguish them from other kinds of events and movements (e.g., Woodward 1998, 2009). Moreover, they map the goal-directed actions of others on to their own actions (Meltzoff 2006). This is an important feature of human causal learning. It distinguishes it from other types of learning, such as classic RL, which only concern an agent's own actions, and also distinguishes it from learning in other animals (Taylor et al. 2014). From early in life, then, children have the cognitive and conceptual structures in place to discover empowerment relations between actions and outcomes, both their own and those of others.

From at least twenty-four months and probably earlier, children make genuine causal inferences by observing the correlations between goal-directed actions—their own or others—and outcomes. However, until around age four, they do not make similar inferences from simple correlations between events (Waismeyer and Meltzoff 2017; Bonawitz et al. 2010; Meltzoff et al. 2012). Suppose a twenty-four-month-old sees a human hand repeatedly push a toy car against a block A, which causes a light to go on. Pushing the car against another block B does not have this effect. Now we ask the infant to make the light appear. Infants will reproduce the correct action on A to make the light go, but not the action on B. However, they will not do this if they simply see the car move on its own and cause the effect. This is true even though they will look toward the light in this condition, suggesting that they have learned the correlation between the motion of the car and the light (Meltzoff et al. 2012). In short, toddlers appear to detect empowerment relations between actions and outcomes and use those relations to infer causal relationships that determine their own future interventions. They do not do this based on correlations among events that do not involve actions and outcomes. Four-year-olds do infer new interventions from correlations alone, but this ability seems to depend on their earlier learning through goal-directed action.

Empowerment also naturally applies to children's early exploratory play (Chu and Schulz 2020). Even infants characteristically play by varying their actions on an object and observing the results—hence the perennial popularity of toys like rattles and busy boxes that afford such empowering actions. Empowerment-based RL, unlike Bayesian inference, also provides a natural way to characterize such experimental actions, they are precisely what you would expect from a system that was trying to act to maximize empowerment.

Thinking of causal learning in terms of empowerment may also help to resolve some of the search problems. Maximizing empowerment would not require the sort

https://doi.org/10.1017/psa.2026.10285 Published online by Cambridge University Press

of search through a high-dimensional hypothesis space that is so challenging for Bayesian inference. Unfortunately, precisely calculating mutual information poses problems of tractability—but some very recent approximation methods make such calculations more feasible (e.g., Zhao et al. 2020). Children might also begin by simply looking for correlations between actions, their own and others, and the outcomes that follow them, rather than fully calculating mutual information.

If children are maximizing empowerment, they would have a mechanism for independently discovering causal relations that are not specified innately, even without requiring the full apparatus of Bayesian causal inference. They might look for relations that have the feature of mutual information between their own actions and those of others and outcomes, like the relations between sticks and toys. This might then allow them to build up a repertoire of basic causal arrows that can then be combined to build more complex models.

In sum, I argue here that recent work on "empowerment" may help bridge Bayesian and RL approaches to learning and provide both empirical and theoretical insight into the crucial problem of learning the causal structure of the world.

Acknowledgments. Discussions with Yu Qing Du, Li Dayan, Eliza Kosoy, Caren Walker, Nihat Ay, Daniel Polani, Pieter Abeel, Charles Kemp, Andrew Perfors, Lerrel Pinto, Yann LeCun, Rosa Cao, James Woodward, and Tom Everitt—along with members of BAIR, the Ranch Metaphysics Workshop, and the Nature of Intelligence Working Group at Santa Fe Institute—all contributed greatly to clarifying these ideas. Maansi Narain was invaluable in helping to prepare the manuscript.

Funding Statement. I would like to thank Google-BAIR Commons, DARPA Machine Common Sense Program, the ONR MURI program, the Templeton Foundation, the Templeton World Charity Foundation, the NSF ARIA AI institute, and CIFAR who funded this research.

Declarations. None to declare

References

Bonawitz, Elizabeth, Stephanie Denison, Alison Gopnik, and Thomas L. Griffiths. 2014a. "Win-Stay, Lose-Sample: A Simple Sequential Algorithm for Approximating Bayesian Inference." *Cognitive Psychology* 74:35–65. https://doi.org/10.1016/j.cogpsych.2014.06.003.

Bonawitz, Elizabeth, Stephanie Denison, Thomas L. Griffiths, and Alison Gopnik. 2014b. "Probabilistic Models, Learning Algorithms, and Response Variability: Sampling in Cognitive Development." *Trends in Cognitive Sciences* 18 (10):497–500. https://doi.org/10.1016/j.tics.2014.06.006.

Bonawitz, Elizabeth, Darlene Ferranti, Rebecca Saxe, Alison Gopnik, Andrew N. Meltzoff, James Woodward, and Laura E. Schulz. 2010. "Just Do It? Investigating the Gap between Prediction and Action in Toddlers' Causal Inferences." *Cognition* 115 (1):104–17. https://doi.org/10.1016/j.cognition.2009.12.001.

Brändle, Fransizka, Lena Stocks, Joshua Tenenbaum, Sam Gershman, and Eric Schulz. 2023. "Empowerment Contributes to Exploration Behaviour in a Creative Video Game." *Nature Human Behavior* 7 (9):1481–89.

Chu, Junyi, and Laura E. Schulz. 2020. "Play, Curiosity, and Cognition." *Annual Review of Developmental Psychology* 2 (1):317–43. https://doi.org/10.1146/annurev-devpsych-070120-014806.

Cohen, Jonathan D, Samuel M McClure, and Angela J Yu. 2007. "Should I Stay or Should I Go? How the Human Brain Manages the Trade-Off between Exploitation and Exploration." *Philosophical Transactions of the Royal Society B: Biological Sciences* 362 (1481):933–42. https://doi.org/10.1098/rstb.2007.2098.

Dayan, Peter, and Bernard W. Balleine. 2002. "Reward, Motivation, and Reinforcement Learning." *Neuron* 36 (2):285–98. https://doi.org/10.1016/s0896-6273(02)00963-7.

De Abril, Ildefons Magran, and Ryota Kanai. 2018. "A Unified Strategy for Implementing Curiosity and Empowerment Driven Reinforcement Learning." Arxiv preprint: 1806.06505.

https://doi.org/10.1017/psa.2026.10285 Published online by Cambridge University Press

Du, Yu-Qing, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, and Anca Dragan. 2020. "Ave: Assistance via Empowerment." *Advances in Neural Information Processing System* 33:4560–71.

Du, Yuqing, Eliza Kosoy, Alyssa Dayan, Maria Rufova, Peter Abbeel, and Alison Gopnik. 2023. "What Can AI Learn from Human Exploration? Intrinsically-Motivated Humans and Agents in Open-World Exploration." *NeurIPS 2023 Workshop: Information-Theoretic Principles in Cognitive Systems*, November. https://neurips.cc/virtual/2023/79197.

Eberhardt, Frederick, and Richard Scheines. 2007. "Interventions and Causal Inference." *Philosophy of Science* 74 (5):981–95. https://doi.org/10.1086/525638.

Giron, Anna P., Simon Ciranka, Eric Schulz, Wouter van den Bos, Azzurra Ruggeri, Björn Meder, and Charley M. Wu. 2023. "Developmental Changes in Exploration Resemble Stochastic Optimization." *Nature Human Behaviour* 7 (11):1955–67. https://doi.org/10.1038/s41562-023-01662-1.

Goddu, Mariel K., and Alison Gopnik. 2024. "The Development of Human Causal Learning and Reasoning." *Nature Reviews Psychology* 3 (5):319–39. https://doi.org/10.1038/s44159-024-00300-5.

Godfrey-Smith, Peter. 2010. "Causal Pluralism." In *The Oxford Handbook of Causation*. Oxford: Oxford University Press, 326–38. https://doi.org/10.1093/oxfordhb/9780199279739.003.0017.

Gopnik, Alison, and Elizabeth Bonawitz. 2015. "Bayesian Models of Child Development." *Wiley Interdisciplinary Reviews: Cognitive Science* 6 (2): 75–86.

Gopnik, Alison, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, and David Danks. 2004. "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets." *Psychological Review* 111 (1):3–32. https://doi.org/10.1037/0033-295x.111.1.3.

Gopnik, Alison, and Henry M. Wellman. 2012. "Reconstructing Constructivism: Causal Models, Bayesian Learning Mechanisms, and the Theory Theory." *Psychological Bulletin* 138 (6):1085–1108. https://doi.org/10.1037/a0028044.

Kidd, Celeste, and Benjamin Y. Hayden. 2015. "The Psychology and Neuroscience of Curiosity." *Neuron* 88 (3):449–60. https://doi.org/10.1016/j.neuron.2015.09.010.

Klyubin, Alexander, Daniel Polani, and Christopher Nehaniv. 2005. "Empowerment: A Universal Agent-Centric Measure of Control." *2005 IEEE Congress on Evolutionary Computation* 1:128–35. https://doi.org/10.1109/cec.2005.1554676.

Leslie, Alan M. 1982. "The Perception of Causality in Infants." *Perception* 11 (2):173–86. https://doi.org/10.1068/p110173.

Meltzoff, Andrew N. 2006. "'Like Me': A Foundation for Social Cognition." *Developmental Science* 10 (1):126–34. https://doi.org/10.1111/j.1467-7687.2007.00574.x.

Meltzoff, Andrew N., Anna Waismeyer, and Alison Gopnik. 2012. "Learning about Causes from People: Observational Causal Learning in 24-Month-Old Infants." *Developmental Psychology* 48 (5):1215–28. https://doi.org/10.1037/a0027440.

Oudeyer, Pierre-Yves, Frdric Kaplan, and Verena V. Hafner. 2007. "Intrinsic Motivation Systems for Autonomous Mental Development." *IEEE Transactions on Evolutionary Computation* 11 (2):265–86. https://doi.org/10.1109/tevc.2006.890271.

Papoušek, Hanuš, and Mechthild Papoušek. 1975. "Cognitive Aspects of Preverbal Social Interaction between Human Infants and Adults." *Novartis Foundation Symposia, January, 241–69.* https://doi.org/10.1002/9780470720158.ch14.

Pathak, Deepak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. "Curiosity-Driven Exploration by Self-Supervised Prediction." *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July, 488–89. https://doi.org/10.1109/cvprw.2017.70.

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

Pearl, Judea. 2000 (2009). *Causality*. 2nd ed. Cambridge: Cambridge University Press.

Piaget, Jean. 1952. *The Origins of Intelligence in Children*. Translated by Margaret Cook. New York: International Universities Press. https://doi.org/10.1037/11494-000.

Rovee-Collier, Carolyn Kent, and Marcy J. Gekoski. 1979. "The Economics of Infancy: A Review of Conjugate Reinforcement." *Advances in Child Development and Behavior* 13:195–255. https://doi.org/10.1016/s0065-2407(08)60348-1.

Salge, Christoph, Cornelius Glackin, and Daniel Polani. 2014. "Empowerment: An Introduction." *Emergence, Complexity and Computation: 67–114.* https://doi.org/10.1007/978-3-642-53734-9_4.

https://doi.org/10.1017/psa.2026.10285 Published online by Cambridge University Press

Schmidhuber, Jürgen. 2010. "Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010)." *IEEE Transactions on Autonomous Mental Development* 2 (3):230–47. https://doi.org/10.1109/tamd.2010.2056368.

Schulz, Laura. 2012. "The Origins of Inquiry: Inductive Inference and Exploration in Early Childhood." *Trends in Cognitive Sciences* 16 (7):382–89. https://doi.org/10.1016/j.tics.2012.06.004.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al. 2018. "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play." *Science* 362 (6419):1140–44. https://doi.org/10.1126/science.aar6404.

Sloan, Aliza, Nancy Jones, and Scott Kelso. 2023. "Meaning from movement and stillness: Signatures of coordination dynamics reveal infant agency." *Proceedings of the National Academy of Sciences* 120 (39): e2306732120.

Spelke, Elizabeth S. 2022. *What Babies Know: Core Knowledge and Composition*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190618247.001.0001.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/1754.001.0001.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Taylor, Alex H., Lucy G. Cheke, Anna Waismeyer, Andrew N. Meltzoff, Rachael Miller, Alison Gopnik, Nicola S. Clayton, and Russell D. Gray. 2014. "Of Babies and Birds: Complex Tool Behaviours Are Not Sufficient for the Evolution of the Ability to Create a Novel Causal Intervention." *Proceedings of the Royal Society B: Biological Sciences* 281 (1787):20140837. https://doi.org/10.1098/rspb.2014.0837.

Ullman, Tomer D., and Joshua B. Tenenbaum. 2020. "Bayesian Models of Conceptual Development: Learning as Building Models of the World." *Annual Review of Developmental Psychology* 2 (1):533–58. https://doi.org/10.1146/annurev-devpsych-121318-084833.

Uzgiris, Ina C., and John M. Hunt. 1975. *Assessment in Infancy: Ordinal Scales of Psychological Development*. Champaign-Urbana: University of Illinois Press.

Waismeyer, Anna, and Andrew Meltzoff. 2017. "Learning to Make Things Happen: Infants' Observational Learning of Social and Physical Causal Events." *Journal of Experimental Child Psychology* 162:58–71. https://doi.org/10.1016/j.jecp.2017.04.018.

Watson, John S. 1972. "Smiling, Cooing and The Game." *Merrill-Palmer Quarterly of Behavior and Development* 18 (4):323–39.

Woodward, Amanda. 1998. "Infants Selectively Encode the Goal Object of an Actor's Reach." *Cognition* 69 (1):1–34. https://doi.org/10.1016/s0010-0277(98)00058-4.

Woodward, Amanda. 2009. "Infants' Grasp of Others' Intentions." *Current Directions in Psychological Science* 18 (1):53–57. https://doi.org/10.1111/j.1467-8721.2009.01605.x.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, James. 2007. "Interventionist Theories of Causation in Psychological Perspective." *Causal Learning*: 19–36. https://doi.org/10.1093/acprof:oso/9780195176803.003.0002.

Woodward, James. 2021. *Causation with a Human Face: Normative Theory and Descriptive Psychology*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780197585412.001.0001.

Yiu, Eunice, Kelsey Allen, Shiry Ginosar, and Alison Gopnik. 2026. "Empowerment Gain and Causal Model Construction: Children and Adults Are Sensitive to Controllability and Variability in Their Causal Interventions." *Philosophical Transactions of the Royal Society A* 384 (2320):20250003. https://doi.org/10.1098/rsta.2025.0003.

Zhao, Ruihan, Pieter Abbeel, and Stas Tiomkin. 2021. "Efficient Online Estimation of Empowerment for Reinforcement Learning." *The Ninth International Conference on Learning Representation*. https://doi.org/10.48550/arXiv.2007.07356.

Cite this article: Gopnik, Alison. 2026. "Empowerment Gain as Causal Learning, Causal Learning as Empowerment Gain: A Bridge between Bayesian Causal Hypothesis Testing and Reinforcement Learning." *Philosophy of Science*. https://doi.org/10.1017/psa.2026.10285

https://doi.org/10.1017/psa.2026.10285 Published online by Cambridge University Press