MIT Press • Open Encyclopedia of Cognitive Science

Mechanistic Explanation

Carl F. Craver[1,2]

[1]Philosophy Department and Philosophy-Neuroscience-Psychology Program,
[2]Washington University in St Louis

MIT Press

Published on: Jan 06, 2025

DOI: https://doi.org/10.21428/e2759450.98d525c7

A *mechanistic explanation* shows how a phenomenon came about or how something works. A distinctive feature of mechanistic explanation is its emphasis on information about the component parts of a system, their activities, and the spatial and temporal constraints on their organization in virtue of which they together produce the system's behavior. Mechanistic explanation is causal, in the broad sense that to explain something causally is a matter of showing how that thing is situated in the causal structure of the world. To explain something is to show what brought it about or to reveal its inner workings. Mechanistic explanation works by showing how causes are arranged in space and time such that they produce or underlie the phenomenon. Among causal explanations, mechanistic explanations are distinctive in revealing internal or intermediate causal structures in addition to distal causes. The request for a mechanism is a request to fill in at least some of the details between cause and effect in order to make it intelligible just how a cause produced the effect or how a capacity works.

History

Both the idea of mechanism and the corresponding notion of mechanistic explanation can be traced across Western intellectual history as a vast, branching lineage following different paths through different areas of science and fitting adaptively into the intellectual niches it occupies. There is no single thing, the mechanical worldview, passed from one generation to the next; neither is there an eternal essence of mechanistic explanation. Instead, there is an evolving lineage of uses that share family resemblances but differ one from the next in key respects (see Boas, 1952; Dijksterhuis, 1961; Gabbey, 2004; Garber, 1992; Roux, 2017). The notion of mechanism has been invoked in many forms (as demanding contact action, determinism, mathematization, geometrical representation, attraction and repulsion, energy transfer) and defended against many kinds of alternatives (souls, vital forces, forms, living substances, emergent powers) over centuries.

The term is now used throughout the special sciences, including biology, neuroscience, and cognitive science. But to fit those sciences, the concept of mechanism has had to stretch and liberalize to accommodate the *mechanisms* found in those domains. What remains of the core mechanistic ideal appears to be only this: We explain things by situating them in the world's causal structure, looking back to their causes and inward to the organized capacities of their component parts [see Causal Learning; Causal Reasoning].

Cognitive science is interesting in this respect because it aims to reveal *cognitive mechanisms*. But mechanisms in that science often do not seem to involve knowing, for example, how brain parts do the things they do. Often they involve laying out a set of computational tasks performed in getting from a presented stimulus to some behavioral indicator but in a way that abstracts from, and idealizes, how the mechanism actually works [see Cognitive Ontology]. To what extent is the goodness of an explanation in cognitive science determined by match with developments in understanding how brains implement the computations in question? And to what extent does a full accounting of the place of minds in nature

require one to look beyond the explanatory resources of an overly narrow conception of the mechanical worldview? Clearly, this is highly disputed territory.

Mechanism has thus become a pregnant and burdened world standing for a set of convictions about the importance of causation, (de)composition, and organization to our ability to render the phenomena in these domains intelligible. These core concepts have been the focus of considerable development in recent philosophy of science, starting with the work of Wesley Salmon (1984) on causal explanation, Herbert Simon (1969) on near decomposability and hierarchy, and William Wimsatt (1997) on forms of organization. These core areas of development are the focus of the next section.

Core concepts

Two kinds of explanatory questions are answered with mechanistic explanations. The first question concerns the causal history (the *etiology*) of a phenomenon: One asks, for example, "How do children develop the capacity for episodic recall?" Here the question takes as the thing to be explained an end-stage of a developmental process. The request or explanation seeks to know the major causal contributors, the arrangement of relevant developmental factors, by which one acquires that form of cognition. Relevant factors might include intrinsic factors, such as the development of the entorhinal cortex and hippocampus, and extrinsic factors, such as parent-child interactions. The etiological explanation allows one to see how the capacity came into existence through the operation of causes arranged in space and time.

A second kind of question asks how a causal system has the capacities, faculties, and processes that it has. For example, How do human beings encode episodic memories? This question asks for a constitutive explanation, a decomposition of the capacity into independent components whose organized activities make it the case that the system encodes episodic memories. And one satisfies this request by beginning to fill in details about the major players (working parts) in this causal system, the things they do, and how they are arranged in space and time so that this system, in fact, has the capacity, faculty, or process in question.

Both questions ask for details about mechanisms. Each situates something puzzling in a network of causes. They differ only in that one looks back to the antecedent causal structures and the other looks within (or downward) to the component parts and organizational features by which the higher-level capacity is explained.

Given the relationships between these questions, three key concepts—causation, composition, and organization—are central to contemporary work on mechanisms.

Causation

The importance of causation to explanation is nicely illustrated with certain *test cases* of bad explanations that any philosophical theory of explanation should be able to diagnose as bad explanations. For example,

the theory should explain why predicting a phenomenon is not the same as explaining it. The length of the shadow and the angle of the elevation of the sun (plus some trigonometry) allow one to predict the unknown height of the flagpole, but those factors do not explain why the flagpole is as tall as it is. Likewise, the falling barometer allows one to predict but not explain the arrival of the storm, but the changes to the barometer are not part of the explanation of the storm. Explanation seems to track causation more than it tracks the ability to make predictions.

Similar examples apply to constitutive explanations. One cannot explain why gases expand when heated by appeal to the ideal gas law precisely because the ideal gas law is merely a general description of the phenomenon and does not say why this phenomena regularity holds (moral: purely phenomenal models are not constitutive explanations of the phenomena they describe). Blood flow increases during cognitive task performance, but that does not mean that changes in blood are explanatory of cognitive task performance (moral: correlations of x to y are not automatically explanations of y by x). Again, it appears that merely deriving a description of the phenomenon to be explained from another set of generalizations is insufficient to explain that phenomenon. And again, it appears that details about the causal organization of the component parts, i.e., about the mechanism, is key to explaining.

Importantly, there is considerable disagreement, even among mechanists, about how to understand causation itself. Some emphasize contact action and mark transmission (e.g., Salmon, 1984). Some ground causation in irreducible activities (Machamer et al., 2000). Others emphasize counterfactual dependence and difference-making (Lewis, 1986) or the invariance of generalizations under intervention (Woodward, 2003; Craver, 2007; Kaplan, 2011). Others rely on regularity theories (Cheng, 1997) or mechanistic accounts (Glennan, 2010). Clearly, what one thinks causation is has a huge influence on what one accepts as a mechanism and what it takes to give a properly mechanistic explanation.

Composition

The second key idea is composition. In constitutive explanations, the phenomenon to be explained is a behavior of a mechanism as a whole, and the explanation describes the component parts and their organization. Our world, especially the world on which evolution has acted, exhibits a hierarchical organization of wholes into nearly decomposable parts (Simon, 1969). Near decomposability, like the idea of *modularity* or *community structure* in network theory, is the thought that components are collections of causally interacting parts that have more or more intense causal interactions with one another than they do with items outside that collection. Interactions with items outside that set are, on that account, interfaces with the other systems (see Haugeland, 1998).

Viewed in this light, the cognitive capacities that are the focus of cognitive scientists appear as high-level capacities or regularities that are to be explained in terms of lower level parts or capacities. They might also be viewed as the components in the explanation of still higher-level capacities. That is, one might be interested in explaining the capacity for episodic memory in humans. In pursuing that, one might be driven to look for mechanisms of encoding, storage, and retrieval of episodic memories. Then, to get

more detailed, one might find differences depending on the kind of information being encoded, e.g., sequential timing versus location, requiring another decomposition into component parts. This kind of iterative decomposition of capacities into sub-capacities gives rise to levels of organization: Each decomposition of a whole into organized collections of parts adds another level of organization to a multilevel mechanism (Craver, 2007). The idea that cognitive systems are organized this way is a kind of organizational principle that unifies different kinds of research on memory as research on different aspects of the same thing; different experimental approaches and models reveal different aspects of a multilevel explanation (Hochstein, 2016).

Organization

Finally, the idea of organization is fundamental to the idea of mechanism. Wimsatt (1997) distinguishes aggregates from mechanisms. In aggregates, the whole is a simple sum of the parts, parts are interchangeable with one another without changing the wholes, adding or subtracting parts changes the feature of the whole linearly, and the spatial and temporal organization of the components is irrelevant to how the parts contribute to a whole. Consider how masses of grains of sand contribute to the mass of the sandpile. In contrast, the parts of a mantle clock are not interchangeable with one another but depend on their spatial organization and temporal coordination. Mechanisms are not mere aggregates but organized, functioning units. Organization, in this sense, involves spatial relations (size, shape, orientation, location, matters of fit), temporal relations (order, rate, duration), and causal relations (feed-forward, feedback, massive connectivity, modularity, small world), all seen as in the service of a capacity to be explained.

Cognitive science often involves the use of experimental effects that reveal features of the causal structure of a mechanism even if the findings do not concern the parts, their physical locations, and the like. The study of mental timing, perhaps the oldest form of cognitive science, revealed causal structures by measuring task durations under various manipulations. Patterns of dissociation and interference are used to argue for different forms of causal independence. We can learn about a mechanism, that is, by studying its higher-level behavior; the behavior itself places constraints on the space of possible mechanisms for explaining it.

Questions, controversies, and new developments

Constitutive relevance

A considerable controversy has arisen over how best to analyze the concept of *constitutive relevance*, i.e., what it is for a part to be a component in a mechanism. Some embrace the idea of mutual manipulability as a sufficient condition on the constitutive relevance relations (see Craver, 2007; Craver et al., 2021; Krickel, 2018a; Krickel, 2018b): If one can intervene on the part and manipulate the whole as well as intervene on the whole to manipulate the part, then one is warranted in thinking the part is a component. Baumgartner and Gebharter (2016) charge that this idea is incoherent and defend an alternative view

according to which a part is constitutively relevant when there exists no intervention on the part relevant to the whole and vice versa (Baumgartner & Cassini, 2017). And there are other approaches as well (Harbecke, 2015; Couch, 2011).

Limits of mechanistic explanation

For many, emphasis on mechanisms is taken to exclude other forms of explanation more appropriate in key domains of biology and cognitive science. Mechanism is, to such critics, an insistence on detail (Batterman & Rice, 2014; Chirimuuta, 2014) or incompatible with dynamical explanation (Chemero & Silberstein, 2008). Others argue that network models, or computational models, offer distinctively non-mechanistic explanatory resources. Cross-cutting this suggestion, some argue that there is a distinctive style of explanation that can appeal to abstract, topological features of a system with relatively sparse attention to its detailed causal structure (Huneman, 2010, 2018). Others posit *teleological explanations*: That is, explanations in terms of the effects of or "purposes for" the phenomenon, which might claim that we have episodic memory because it gives us as a way of coping with the limits of our rule-based, semantic generalizations about people and circumstances (Klein et al., 2002). Others have argued that intentional explanations in terms of goals and intentions go beyond mere understanding of how something is situated in the causal order and additionally includes how the states are situated in a rational order (Dennett, 1989). Finally, some posit the existence of distinctively mathematical explanations (Lange, 2013; Povich, in press).

Such discussions hinge on how one characterizes the limits of mechanistic explanation and upon whether these other forms of explanation are, in fact, noncausal in the relevant sense, once one understands how they function. Many mechanists hope to encompass teleological and intentional explanations within a broadly causal-mechanical view of explanation, although such proposals remain contested and concern the very limits of causal-mechanical worldview for understanding minds. Mathematical and teleological explanations on the other hand, might reflect deeper disagreement about what it means to explain something.

Varieties of mechanistic explanation

An active area of development concerns the effort to catalog different types of causal-mechanical explanations, different ways of organizing components into collective activity (Craver & Darden, 2013; Glennan, 2017; Ross, 2021). The idea is that different ways of arranging components (e.g., closed, open, feed forward, feed-back, cascades) have implications for the aims of discovering, understanding, and controlling mechanisms (see Glennan et al., 2022).

Simplicity of mechanism

There is also a further, and related, question about whether the mechanistic approach to explanation is overly simplistic, that it fails to recognize kinds of complexity that must be acknowledged in a full accounting of phenomena. The idea of a mechanism seems to suggest to many a stepwise machine, with

localized parts performing largely dissociable functions. But if one looks at any dynamical system, or even a common circuit board, that mode of understanding (knowing the parts and how they are arranged causally with respect to one another) fails to capture the key dynamics of the interaction of the parts [see Complex Dynamical Systems]. Just how the mode of understanding deployed in understanding circuit board differs from, or contrasts with, mechanistic explanation is a topic of future development.

Broader connections

Work on mechanisms in the philosophy of science makes considerable contact with work on emergence, reduction, and higher-level causation. In general, contemporary mechanists have presented their work, emphasizing parts and wholes and levels, as an alternative to reduction for thinking about the structure of science. Higher-level causal relationships, such as those typically described in models of cognitive processes, in this view, are underwritten by mechanisms and so not at all mysterious. Explanations across levels are required to integrate work across many areas of science (from ions and molecules to organisms and societies). Higher-level causes in this sense are typically contrasted with *spooky* forms of emergence, which seem to involve wholes having properties or activities that cannot be mechanistically explained.

The aim of cognitive science, from a mechanistic perspective, is to show how different forms of cognition are possible in a world of causes, to situate the mind in the causal structure of the world. Marr (1982), for example, emphasized different descriptions of one and the same mechanism: one emphasizing the mathematical function being computed, one describing the specific algorithm by which that abstract problem is solved, and the third emphasizing its hardware implementation, the actual nuts and bolts by which the task is carried out. Marr insisted on an independence of the computational level from details about the implementation. Others insist oppositely on the need to consider lower-level mechanisms in evaluating computational explanations (Piccinini, 2020). Clearly, the question turns in part on whether there are elements in our understanding of the mind that resist capture within the causal language of the mechanist and the forms of explanation they accept.

Although philosophical work on mechanism is often contrasted with work on causal modeling, work across that sociological divide promised to bear important fruit as computational approaches take center stage in science. Most causal modeling frameworks (e.g., Pearl, 2009; Spirtes et al., 2000; Woodward, 2003) work at a single level, in marked contrast to the importance of interlevel explanations in sciences such as cognitive science and neuroscience. Philosophical work on mechanisms might help to frame problems to be addressed by more formal discovery methods. One particularly fruitful place for research concerns the relationship between formal notions of causal mediation and mechanistic notions of *constitutive relevance*.

Finally, philosophical work on mechanism has thus far had little interaction with work on the psychology of causal inference and the nature of mechanistic reasoning (see Bachtiar, in press). How does reasoning about mechanisms interact with reasoning about counterfactual relations in the assessment of explanatory generalizations? How do humans track not just matters of causal relevance but matters of

mechanistic componency? How are mechanistic representations implemented cognitively? How do infants and adults reason about mechanisms? How can mechanistic reasoning be taught? Research on the development of causal learning might be extended to investigate the causal mechanisms by virtue of which we understand the mechanisms by which those causes deliver their effects.

Further reading

- Bechtel, W. (2007). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Psychology Press. https://doi.org/10.4324/9780203810095
- Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441. https://doi.org/10.1016/j.shpsc.2005.03.010
- Craver, C. F. (2007). *Explaining the brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199299317.001.0001
- Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press. https://doi.org/10.1093/oso/9780198779711.001.0001

References

- Bachtiar, R. W. (in press). *Understanding phenomena: Mechanistic reasoning in science education.* Routledge. ↩
- Batterman, R. W., & Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, *81*(3), 349–376. https://doi.org/10.1086/676677 ↩
- Baumgartner, M., & Casini, L. (2017). An abductive theory of constitution. *Philosophy of Science*, *84*(2), 214–233. https://doi.org/10.1086/690716 ↩
- Baumgartner, M., & Gebharter, A. (2016). Constitutive relevance, mutual manipulability, and fat-handedness. *The British Journal for the Philosophy of Science*, *67*(3), 731–756. https://doi.org/10.1093/bjps/axv003 ↩
- Boas, M. (1952). The establishment of the mechanical philosophy. *Osiris*, *10*, 412–541. https://doi.org/10.1086/368562 ↩
- Chemero, A., & Silberstein, M. (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science*, *75*(1), 1–27. https://doi.org/10.1086/587820 ↩

- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405. https://doi.org/10.1037/0033-295X.104.2.367

  ↩
- Chirimuuta, M. (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*, *191*(2), 127–153. https://doi.org/10.1007/s11229-013-0369-y

  ↩
- Couch, M. B. (2011). Mechanisms and constitutive relevance. *Synthese*, *183*(3), 375–388. https://doi.org/10.1007/s11229-011-9882-z

  ↩
- Craver, C. F. (2007). *Explaining the brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199299317.001.0001

  ↩
- Craver, C., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.

  ↩
- Craver, C., Glennan, S., & Povich, M. (2021). Constitutive relevance & mutual manipulability revisited. *Synthese*, *199*(3), 8807–8828. https://doi.org/10.1007/s11229-021-03183-8

  ↩
- Dennett, D. C. (1989). *The intentional stance*. MIT Press.

  ↩
- Dijksterhuis, E. J. (1961). The mechanization of the world picture. *Science and Society*, *35*(2), 232–238.

  ↩
- Gabbey, A. (2004). What was "mechanical" about "the mechanical philosophy"? In C. R. Palmerino & J. M. M. H. Thijssen (Eds.), *The reception of the Galilean science of motion in seventeenth-century Europe* (pp. 11–23). Springer. https://doi.org/10.1007/978-1-4020-2455-9_2

  ↩
- Garber, D. (1992). *Descartes' metaphysical physics*. University of Chicago Press.

  ↩
- Glennan, S. (2010). Mechanisms, causes, and the layered model of the world. *Philosophy and Phenomenological Research*, *81*(2), 362–381. https://doi.org/10.1111/j.1933-1592.2010.00375.x

  ↩

- Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press. https://doi.org/10.1093/oso/9780198779711.001.0001

↩

- Glennan, S., Illari, P., & Weber, E. (2022). Six theses on mechanisms and mechanistic science. *Journal for General Philosophy of Science*, *53*(2), 143–161. https://doi.org/10.1007/s10838-021-09587-x

↩

- Harbecke, J. (2015). The regularity theory of mechanistic constitution and a methodology for constitutive inference. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *54*, 10–19. https://doi.org/10.1016/j.shpsc.2015.09.004

↩

- Haugeland, J. (1998). *Having thought*: *Essays in the metaphysics of mind*. Harvard University Press.

↩

- Hochstein, E. (2016). One mechanism, many models: A distributed theory of mechanistic explanation. *Synthese*, *193*(5), 1387–1407. https://doi.org/10.1007/s11229-015-0844-8

↩

- Huneman, P. (2010). Topological explanations and robustness in biological sciences. *Synthese*, *177*(2), 213–245. https://doi.org/10.1007/s11229-010-9842-z

↩

- Huneman, P. (2018). Diversifying the picture of explanations in biological sciences: Ways of combining topology with mechanisms. *Synthese*, *195*(1), 115–146. https://doi.org/10.1007/s11229-015-0808-z

↩

- Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183*(3), 339–373. https://doi.org/10.1007/s11229-011-9970-0

↩

- Klein, S. B., Cosmides, L., Tooby, J., & Chance, S. (2002). Decisions and the evolution of memory: Multiple systems, multiple functions. *Psychological Review*, *109*(2), 306–329. https://doi.org/10.1037/0033-295X.109.2.306

↩

- Krickel, B. (2018a). *The mechanical world: The metaphysical commitments of the new mechanistic approach*. Springer.

↩

- Krickel, B. (2018b). Saving the mutual manipulability account of constitutive relevance. *Studies in History and Philosophy of Science Part A*, *68*, 58–67.

↩

- Lange, M. (2013). What makes a scientific explanation distinctively mathematical? *The British Journal for the Philosophy of Science*, *64*(3), 485–511. https://doi.org/10.1093/bjps/axs012

↩

- Lewis, D. (1986). Causal explanation. In D. Lewis (Ed.), *Philosophical Papers* (Vol. II, pp. 214–240). Oxford University Press.

↩

- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.1086/392759

↩

- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* W. H. Freeman.

↩

- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

↩

- Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press. https://doi.org/10.1093/oso/9780198866282.001.0001

↩

- Povich, M. (in press). *Rules to infinity: The normative role of mathematics in scientific explanation*. Oxford University Press.

↩

- Ross, L. N. (2021). Causal concepts in biology: How pathways differ from mechanisms and why it matters. *The British Journal for the Philosophy of Science*, *72*(1), 131–158. https://doi.org/10.1093/bjps/axy078

↩

- Roux, S. (2017). From the mechanical philosophy to early modern mechanisms. In S. Glennan & P. Illari (Eds.), *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 26–45). Routledge. https://doi.org/10.4324/9781315731544-3

↩

- Salmon, W. (1984). Scientific explanation and the causal structure of the world. Princeton University Press.

  ↩
- Simon, H. A. (1969). *The sciences of the artificial*. MIT Press.

  ↩
- Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). MIT Press.

  ↩
- Wimsatt, W. C. (1997). Aggregativity: Reductive heuristics for finding emergence. *Philosophy of Science*, *64*(S4), S372–S384. https://doi.org/10.1086/392615

  ↩
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press. https://doi.org/10.1093/0195155270.001.0001

  ↩