Biol Philos (2009) 24:137–145
DOI 10.1007/s10539-008-9123-0

# Carl F. Craver, Explaining what? Review of explaining the brain: mechanisms and the mosaic unity of neuroscience

## Clarendon Press–Oxford University Press, 2007, 272 pp, $49.50 (06)

**Arnon Levy**

Published online: 6 August 2008


**Abstract** Carl Craver's recent book offers an account of the explanatory and theoretical structure of neuroscience. It depicts it as centered around the idea of achieving mechanistic understanding, i.e., obtaining knowledge of how a set of underlying components interacts to produce a given function of the brain. Its core account of mechanistic explanation and relevance is causal-manipulationist in spirit, and offers substantial insight into casual explanation in brain science and the associated notion of levels of explanation. However, the focus on mechanistic explanation leaves some open questions regarding the role of computation and cognition.

**Keywords** Explanation · Neuroscience · Mechanisms · Explanatory relevance

As the name suggests, *Explaining the Brain* is a book about explanation in neuroscience. Carl Craver, who has argued in recent years for the importance of mechanistic explanation in biology, portrays contemporary brain science as centered around the goal of uncovering the mechanistic underpinnings of brain functions.

The introduction states three central desiderata for an account of explanation. The first is descriptive adequacy—a philosophical account of explanation should seek to capture the complexity of real-life scientific explanations. The second is demarcation—an account of explanation should explain what is special about it. The third is normativity—the account should allow us to assess explanations as good or bad, better or worse. To a large extent, these desiderata shape the argumentative style of this book. Craver stays close to actual neuroscientific explanations, and an

A. Levy (✉)
Harvard University, Cambridge, MA, USA
e-mail: levy3@fas.harvard.edu

insistence on the primacy of mechanistic explanation implies clear criteria for demarcation and assessment.

Thematically, the book is divided into three parts. The first two chapters are introductory—beginning with a statement of the book's main goals and a summary of its important theses. Craver then reviews some central arguments from the literature on explanation that buttress the causal view of explanation. These should be familiar to most people in the field. The second part, no doubt the heart of the book, includes chapters 4 and 5 and contains an account of causal-mechanistic explanation. The third part discusses the overall theoretical structure of neuroscience, given the mechanistic framework argued for in earlier chapters.

In many ways, this is an impressive piece of work. It is rich and comprehensive. It manages to advance a number of substantive philosophical theses while staying sensitive to the content and the history of the science in question. Craver offers insightful accounts of three central philosophical issues that arise in understanding the theoretical structure of neuroscience: the nature of explanatory relevance in neuroscience, the role of reduction, and the question of what unifies the various disciplines in which the brain is nowadays studied.

But the book also suffers from significant shortcomings. These have to do primarily with the scope of the discussion and consequently with the overall picture of brain science it presents. On the one hand, it is difficult to see why the mechanistic account on offer applies uniquely, or even especially to the brain. On the other hand, reaching the end of the book, one feels that an important part of what is uniquely interesting about neuroscience has been left out.

## Mechanistic explanation and the structure of neuroscience

Craver remains relatively coy about the question: "what is a mechanism?" In line with his earlier views on the topic (Machmar et al. 2000; Craver 2001) he describes mechanisms as sets of entities with associated activities, organized so as to "exhibit" or "constitute" some specific phenomenon. His liberal reading of "entities" and their associated "activities" implies that pretty much any organized set of causal components constitutes a mechanism. This seems like the right attitude to take: "mechanism" is an umbrella term for an extended family of explanatory structures. The important contrast, highlighted along the way, is between a mechanism and an etiology—the former is the causal structure *underlying*, or *constituting* a phenomenon, while the latter is the causal sequence *leading up to it*. Both serve in explanation, of course, but of different sorts.

Craver operates with several central examples of mechanistic explanation in neuroscience; perhaps the one he refers to most often is the mechanism of long-term potentiation (LTP). A brief summary of LTP should give the flavor of the sorts of explanations Craver focuses on. LTP is a much-studied form of synaptic plasticity, in which co-activation of pre- and post-synaptic neuron induces a persistent increase in the subsequent efficiency of a synapse. LTP is a form of conditioning, in which "successful" activation results in increased sensitivity to the activator. For this reason, many view it as a potentially central mechanism underlying learning and

memory (though the evidence for its centrality is relatively limited). Much is known about the cellular mechanisms underlying LTP.[1] The gist of it is as follows. Activity on the pre-synaptic side induces a receptor on the post-synaptic side, the NMDA receptor, to change conformation so as to become a calcium ($Ca^{2+}$) channel. But, unless the post-synaptic side is also active this calcium channel remains blocked by magnesium ions ($Mg^{2+}$, carrying the same charge as $Ca^{2+}$). Upon depolarization, the magnesium block is released,[2] and calcium flows into the cell. The flow of calcium triggers a number of biochemical pathways resulting, in the short term, in an increase in the number of post-synaptic receptors and in the long-term generating a non-calcium dependent increase in sensitivity. Thus, the post-synaptic cell is effectively equipped with a coincidence detector, sensing a "success" and increasing synaptic strength as a response to it. Craver, like many brain scientists, takes this account of LTP to be a paradigm success story of explanation in neuroscience.

Chapter 2 offers a concise review of arguments for the causal approach to explanation. Craver then sets out to defend a manipulability account of causal explanation, in the style of Woodward (2003). On this view, a relevant causal link exists between two variables, *X* and *Y*, just in case one can (in principle, at least) intervene on *X* in order to change *Y*. Roughly speaking, an intervention on *X* is a change in the value of *X* that changes *Y* only via the change in *X* (i.e., neither directly affecting *Y*, nor via a common cause). The discussion of the manipulability approach follows closely in Woodward's footsteps, with LTP as an illustration. Necessary but fairly familiar ground is covered in this chapter, and an impatient reader might feel that Craver could have been terser. The manipulability account is not short on illustrative examples, and LTP, though central to neuroscience, does not shed much new light on manipulability as such. But the discussion is helpful in setting the stage for Craver's own contributions.

Chief among these contributions is a manipulability-based account of mechanistic explanation. It marries Woodwardian constrains on causal description to a focus on constitutive, analytical explanation. Craver takes mechanistic explanation to be a species of analytical explanation: the breaking down of a complex phenomenon into ingredients (as discussed by philosophers of psychology, e.g., Cummins 1975, 1983; Dennett, 1983; Haugland 1998). He seeks an account that goes beyond describing the strategy of mechanistic explanation, instead delineating (in line with the second initial desideratum) the set of norms, or regulative ideas, which brain scientists employ when setting themselves explanatory goals, and according to which they evaluate work in the field. Such goals, fully formulated, should allow one to evaluate mechanistic explanation along two key dimensions:

[1] LTP is mediated through a large number of molecular mechanisms in the brain, depending on cell-type, region, age, and other factors. Here I am describing only the so-called NMDA-receptor dependent LTP, perhaps the best-understood LTP mechanism, and the one Craver discusses too.

[2] In truth, these are all stochastic processes. The block is not released, like a plug. Rather, the mean time spent by magnesium ions inside the channel decreases, and correspondingly, the mean time in which calcium can travel inward increases. It is important to bear in mind that mechanistic descriptions of the sort offered in biology textbooks, and discussed by philosophers, are very often stylized versions of the probabilistic soup actually in place.

1. *Factivity* They would distinguish a truly explanatory model from a potentially explanatory one (in Craver's terms, they distinguish *how-possibly* explanations from *how-actually* explanations).
2. *Completeness* They would tell us what makes for completeness in mechanistic explanation (what distinguishes mechanism *sketches* as opposed to complete mechanisms).

[The labels—factivity and completeness—are mine].

Craver narrows the focus of his approach a little, relative to the analytical approach of Cummins, for instance. He does not seek an account that includes non-constitutive analytical explanations, those that account for a capacity of a system in terms of other capacities it have (as when the chef's cooking is explained in terms of his ability to read recipes, or to dice vegetables). As a result, he can operate with a relatively simple notion of component: a component is a spatiotemporal sub-part of a system. Of course, not any sub-part will do, and Craver is not able, nor willing, to provide a further characterization that would tell us which spatiotemporal parts are a system's real components (the 'real' is crucial, if one to is to meet (1)). I think he rightly considers this to be a matter for which no general criteria can be given, and on which little confusion is likely to arise in interesting cases.

One issue which Craver discusses here, but does not, one feels, pay enough attention to, is the contrast between mechanistic and aggregative explanations. The latter are found in statistical physics and in population biology, for instance, where the properties of an ensemble of individuals—diffusive fluxes, changes in gene frequency and suchlike—are explained in terms of the properties of large aggregates of similar individuals. No doubt this is a real and important contrast, a serious discussion of which lies outside the scope of *Explaining the Brain*. But Craver gives readers the sense that the distinction between the categories is straightforward: aggregates are merely the sum of their parts, whereas in mechanisms *organization* is key (he cites conditions from Wimsatt (1997) to this effect). But the ground here is far from cleared. One concern is that there are many important intermediate cases, in which organization is to some extent statistical, and in which more than the mere distinction between the two is needed if one is to illuminate the explanatory goings-on. Another is that in many instances of mechanistic organization the aggregative properties of the components are crucial to understanding the activities in question, such that without them the mechanistic account is largely unilluminating. Neuro-electrophysiology provides important examples. The ionic currents that form components in a mechanistic description of, say, action potentials and LTP are aggregative phenomena, and the organization of the mechanism stems in part from the statistical properties of the ensembles of molecules involved. (Another fascinating example in this context is neural networks, but these—oddly—are hardly mentioned in the book.)

Craver next presents his main contribution to the literature on mechanistic explanation—an account of constitutive relevance. Causal theories of explanation have typically had trouble accounting for explanatory relevance: whereas it is fairly clear that much, if not all, explanation in science is causal, it is also clear that many parts of the etiology, or (more relevantly in the present context) the underlying

mechanism are typically left out of the best explanation on offer. We do not mention that gravity is pulling on the brain when accounting for LTP, nor do we, typically, make reference to the evolutionary history of pyramidal cells. Reviewing and classifying a variety of experiments through which candidate mechanisms are tested—e.g., stimulation, interference, and compensation—Craver suggests that constitutive relevance consists in *mutual* manipulability. That is, a component $\varphi$ is relevant to workings of a capacity $\psi$ just in case it is possible to intervene on $\varphi$ so as to change $\psi$, and vice versa. In the example of LTP it is possible, say, to block calcium influx, and thereby inhibit long-term changes in synapse sensitivity; and one can intervene to increase, the likelihood and magnitude of LTP by providing the cell with a train of stimuli (a so-called "tetanus"), thereby also increasing calcium influx. Note, of course, that the requirement is for there to be at least one possible intervention in each direction, not that every intervention on $\varphi$ or $\psi$ to be of this sort.

The discussion of mechanistic relevance is one of the main achievements of the book. Though some of the details of the account might be contested, Craver offers a tight set of conditions and offers an account of explanatory ideals that is acutely sensitive to the manner in which the explanations in question are constructed, tested, and evaluated. The mutual manipulability requirement captures nicely the synergy of top-down and bottom-up experiments in brain science (and other mechanistic disciplines), especially in accounts of mechanisms underlying behavioral capacities, where one manipulates the behavioral tasks set to an organism and attempts to observe underlying changes (e.g., fMRI, or in direct recording of neural activity in laboratory animals), or (more frequently) when cell-level structures are manipulated, genetically or otherwise, and effects on higher capacities are monitored. Mutual manipulability embeds mechanistic explanation within a more general approach to causal–analytical explanation, a project which, surprisingly, has rarely been attempted in the literature on these topics.

One cause for concern regarding this discussion ought to be mentioned. It relates to the factivity desideratum. It is now fairly widely accepted amongst philosophers and to some extent by active scientists as well, that many explanatory models are not, in fact, factive. At least, not in their apparent content. For instance, many explanatory models contain idealizations, elements that are known to be false but nevertheless contribute, often irreplaceably, to the explanatory power of a model. Thus, explaining actual phenomena often involves appeal to fictions. Neuroscience is rife with idealized models, from the Hodgkin-Huxley model of the action potential[3]—in which the neuron is taken to be cylindrical and various non-uniformities (in the conductance properties of the membrane, e.g.) are "smoothed

[3] Action potentials are fleeting voltage changes that travel along axons, AKA nerve "firings". Hodgkin and Huxley won the Nobel prize in 1963 for elucidating the "ionic mechanisms involved in excitation and inhibition…of the nerve cell membrane", as the Nobel Prize committee put it. Craver argues that the Hodgkin-Huxley model is not, in fact explanatory. He views it as an empirical (mathematical) description that does not elucidate the mechanism underlying the action potential. I believe this argument is mistaken as a view of the historical achievement of Hodgkin and Huxley, and possibly even as an application of Craver's own philosophical views about explanation. But this is not the appropriate context for discussing this issue thoroughly.

out"—through network and computational models. Craver describes the condition I have labeled factivity as the requirement that an account of explanation distinguish "loosely constrained conjectures" from "real components, activities, and organizational features" (p. 112), and describes these as two ends of a spectrum on which lie various "*how-plausibly* models", models which are "more or less consistent with known constraints on the components, their activities, and their organization" (pp. 121–122). This description leaves no room for idealization, a deliberate construction of models inconsistent with "known constraints". It is not unlikely that our best understanding of some aspects of the brain, especially if that understanding is to be quantitative, will involve prominent idealizations. Some of the most successful models currently on offer do. A theory of the norms of explanation in neuroscience must, at the very least, acknowledge the role idealization. It would be best to have an account that naturally encompassed both idealized and non-idealized models.

The final portion of Explaining the Brain—chapters 5, 6, and 7—uses the earlier account of mechanisms to draw a big-picture view of the theoretical structure of neuroscience. It depicts neuroscience as engaged in the elucidation of a nested hierarchy of mechanisms in which functions at one level serve as components in a mechanism performing functions at a higher level.

Chapter 5 starts out by distinguishing various senses of levels and arguing that the relevant sense in which explanations in neuroscience span multiple levels is the constitutive sense: the lower levels entities are components in mechanisms at the higher level. Craver rejects several existing criteria for individuating levels: mereological levels, levels of aggregation (neither very substantial candidates to begin with), as well as spatial containment, a most natural, but insufficient criterion. He duly acknowledges the pragmatic, context-dependent nature of divisions into levels, which depends on a prior carving-up of the phenomena into explananda driven by theoretical interests.

Chapter 6 discusses the causal-exclusion argument, best-known from work of Kim (1989, 1993) which might, Carver apparently fears, be used to oppose the idea that non-fundamental levels can figure in explanations. This fear seems understantiated. The exclusion argument is made in the context of metaphysical discussions of, roughly speaking, the mind–body relation. It is supposed to show (again, roughly speaking) that if one accepts that the mental supervenes on the physical, then one ought to view the mental as epiphenomenal, as causally inert. The mechanism–component relation is not a supervenience relation and, thus, the argument from exclusion does not bear on the efficacy or the explanatory relevance of non-fundamental levels in the mechanistic sense. This is pretty much the answer that Craver gives as well, and, as he notes, Kim himself makes the point in his original formulation of the argument (p. 211). In light of this, it is odd that upwards of thirty pages are spent on the matter.

The closing chapter is devoted to putting pieces together. I found it the most enjoyable to read. In it Craver draws a broad brush-stroke picture of the structure of explanation in neuroscience, and illustrates (not for the first time, but in an incisive way) how some of the motivation for the mechanisms approach lies in its offering a systematic alternative to reduction as a view of the bedrock, regulative goal of natural science. While saying exactly what reduction amounts to is not an easy task,

the contrast between a reductionist attitude and a mechanistic one is vivid enough. The reductionist seeks to get rid of apparent, higher-level phenomena whereas for the mechanist they are (in an explanatory context, at least) what gives lower-level entities their identity: entities and activities at lower levels are what they are by virtue of being components in a mechanism. Craver traces some of the history of research on memory, revisiting, in part, the case of LTP, and the hierarchy of memory mechanisms in which it is embedded. He seeks to show that it is mistake to think of neuroscience—or at least the study of memory and learning—as exhibiting a reductionist historical trend. It is not fully clear what sort of direct support this historical point, correct as it may be, lends to the philosophical thesis in question; but it does serve to undercut arguments that celebrate a reductionist trend. That neuroscience is at least in part in the business of constructing mechanistic hierarchies of this sort is, I think, convincing and illuminating.

## Scope and motivation

Why write a book about explanation in neuroscience? There can be two answers to this question, niether exclusive of the other. Explanation in brain science might be interesting in its own right, and it might serve as an example (perhaps a paradigmatic example) of explanation in some broader area of science (perhaps science as a whole). Which of these goals did Craver have in mind in writing *Explaining the Brain?* In the very beginning he states that his aim is to develop "a unified framework for the philosophy of neuroscience" and that "because neuroscience is like other special sciences in many respects, this framework carries lessons for the philosophy of science generally" (p. vii). Thus, the intention is to show how neuroscience is a distinct area of study, but also provide lessons of a more general sort. I had concerns as to whether the book makes good on this two-pronged promise. On the one hand, although it touches frequently on cell biology, genetics, structural biology, and related disciplines, these are only discussed to the extent that they deal with parts of the brain and with neural processes. And, as we've seen, considerable space is devoted to the overall structure of neuroscience, suggesting that mechanistic unity accounts for what is distinct about neuroscience. General lessons are mostly left implicit. On the other hand, some generally agreed-upon ways in which the brain is unique are left out entirely: the book hardly discusses cognition and, specifically, does not mention computational explanation in brain science.

This raises several questions. The kind of mechanistic explanations Craver focuses on are ubiquitous in cell biology, genetics, and physiology and one worries that targeting their role in neuroscience obscures aspects of explanation in biology that become apparent when one looks at mechanistic explanation across the biological sciences. The role of aggregative explanation, noted above, is one such example. Furthermore, emphasizing the role of mechanisms akin to LTP in neuroscience tends to obscure their significance in other contexts and their connections to other forms of explanation. So it might seem that the exclusive focus on mechanisms in neuroscience obscures contrasts and likenesses between

neuroscience and other parts of biology, and makes the drawing of general lessons difficult.

On the other hand, the absence of cognition is rather startling. The brain is, as the title of Paul Churchland's book goes "the engine of reason [and] seat of the soul". And there is a large and increasingly well-developed program of explaining cognitive capacities in computational terms. Many, if not most, of the people who study the brain, including many of the biologists and the philosophers discussed in the book, believe that it is, in essence, a device for processing information, and that explaining how it carries out cognitive functions will eventually consist, to a significant extent, in elucidating the various computational routines performed by the brain, as well as its basic underlying formal architecture. This sort of endeavor is connected, of course, to an understanding of the cellular and molecular mechanisms underpinning cognition—among them the action potential, LTP and other examples discussed by Craver. But there is (widely believed to be) a kind of explanation of these functions that is distinct from the biochemical and cell-biological ones, an explanation couched in the language of inputs, algorithms, and outputs. In a sense, computational descriptions are supposed to form a link between the lower-level mechanisms that Craver is mostly focused on, and the mental processes which these mechanisms give rise to.[4] Indeed some take this to be their main attraction.

Computational explanation does not seem to be mechanistic. While it is likely a form of constitutive or analytical explanation, one that breaks down an explanadum into constituents, these constituents are not spatiotemporal subparts of anything. More generally, it is unclear that the criteria for explanatory relevance defended in the book are necessary or sufficient to encompass computational explanation.

Perhaps Craver holds that computational neuroscience is not as promising as it is advertised to be. Or perhaps he believes that it ought to be discussed separately. He argues for neither view.

To those that believe that biology in general is organized around the goal of supplying mechanistic understanding, this work can serve as the beginning, but surely not the end of an argument for such a conception. Mechanistic explanation is prominent in cell biology and related disciplines, and something akin to Craver's account might fit this wider area of biology nicely. But aggregative explanations figure prominently in population biology and in evolutionary biology and in some areas of physiology, and there are important types of non-mechanistic analytical explanation in areas such as computational neuroscience. Given this variety, one would need to make substantial amendments to the mechanistic outlook if it is to serve as a general view of explanation in biology, describing an ideal to which biological understanding aspires. Whatever one's view of this broader issue, Craver's silence on it means that his conception of neuroscience bears an uncertain relation to the wider scientific context in which brain science is embedded.

[4] For a recent attempt to spell out what computational explanation consists in, giving close attention to neuroscientific exemplars, see Shagrir (2006)

That said, there is plenty of subtle philosophy in *Explaining the Brain*, and much of the details, as well as the underlying spirit of a mechanistic outlook will surely be retained in a more inclusive account of explanation in biology.

## References

Craver CF (2001) Role functions, mechanisms and hierarchy. Philos Sci 68:31–55. doi:10.1086/392866
Cummins R (1975) Functional analysis. J Philos 72:741–746. doi:10.2307/2024640
Cummins R (1983) The nature of psychological explanation. Bradford/MIT Press, Cambridge, MA
Dennett D (1983) The intentional stance. Bradford/MIT Press, Cambridge, MA
Haugland J (1998) Having thought. Harvard University Press, Cambridge, MA
Kim J (1989) Mechanism, purpose and explanatory exclusion. Philos Perspect 3:77–108. doi:10.2307/2214264
Kim J (1993) Supervenience and mind. MIT Press, Cambridge, MA
Machmar P, Darden L, Craver C (2000) Thinking about mechanisms. Philos Sci 67:1–25
Shagrir O (2006) Why we view the brain as a computer. Synthese 153:393–416
Wimsatt W (1997) Aggregativity: reductive heuristics for finding emergence. In: Reengineering philosophy, Philosophy of Science, pp S372–S384
Woodward J (2003) Making things happen: a theory of causal explanation. Oxford University Press, Oxford