Explaining the Brain: Ruthless Reductionism or Multilevel Mechanisms?[1]

Markus Eronen
University of Osnabrück
Institute of Cognitive Science
maeronen@uos.de

Abstract: Mechanistic explanation and metascientific reductionism are two recent and widely discussed approaches to explanation and reduction in neuroscience. I will argue that these are incompatible and that mechanistic explanation has a stronger case, especially when it is combined with James Woodward's manipulationist model of causal explanation.

1. Introduction

In this paper, I will compare and criticize two approaches to reduction and explanation in neuroscience: metascientific reductionism and mechanistic explanation. I will first show that the traditional models of intertheoretic reduction are unsuitable for neuroscience. Then I will compare John Bickle's model of metascientific reductionism and Carl Craver's model of mechanistic explanation, arguing that the latter has a stronger case, especially when supplemented with James Woodward's interventionist account of causal explanation.

2. Intertheoretic reduction

The development of intertheoretic models of reduction started in the middle of the 20th century, in the spirit of logical positivism. The ultimate goal was to show how unity of science could be attained through reductions. In the classic model (most importantly Nagel 1961, 336-397), reduction consists in the deduction of a theory to be reduced (T_2) from a more fundamental theory (T_1). Conditions for a successful reduction are that (1) we can connect the terms of T_2 with the terms T_1, and that (2) with the help of these connecting assumptions we can derive all the laws of T_2 from T_1.

[1] This is the final draft of a paper that was presented at the 31st International Wittgenstein Symposium in Kirchberg am Wechsel, Austria, 10.-16.8.2008. Thanks to Achim Stephan and Jani Raerinne for comments.

Unfortunately this model fails to account for many cases that are regarded as reductions. The model is too demanding: it is very hard to find a pair of theories that would meet these requirements. Even Nagel's prime example, the reduction of thermodynamics to statistical mechanics, is much more complicated than Nagel thought (see, *e.g.*, Richardson 2007). The classic model also has problems accommodating the fact that the reducing theory often *corrects* the theory to be reduced, which means that the theory to be reduced is strictly speaking false. However, logical deduction is truth-preserving, so it should not be possible to deduce a false theory from a true one.

Problems of this kind lead to the development of more and more sophisticated models of intertheoretic reduction, and finally to the "New Wave reductionism" of P. S. Churchland (1986), P. M. Churchland (1989) and J. Bickle (1998, 2003, 2006). Due to constraints of space, I will not go through these models here. It is sufficient to point out one fundamental assumption that underlies *all* intertheoretic models of reduction, and which leads to serious problems in the case of psychology and neuroscience.

This assumption is that the relata of reductions are exclusively *theories*, and that inter*theoretic* relations are the only epistemically and ontologically significant interscientific relations (see, *e.g.*, McCauley 2007). However, well-structured theories that could be handled with logical tools are rare in and peripheral to psychology and neuroscience. Instead, scientists typically look for mechanisms as explanations for patterns, effects, capacities, phenomena, and so on (see, *e.g.*, Machamer *et al.* 2000 and Cummins 2000). Although there are theories in a loose sense in psychology and neuroscience, like the LTP theory for spatial memory or the global workspace theory, these are not theories that could be formalized, and can hardly be the starting points or results of logical deductions. Therefore looking at the relations between theories is the wrong starting point, at least in the case of psychology and neuroscience.

3. Metascientific reductionism

At least partly for these reasons, John Bickle, the most ardent advocate of New Wave reductionism, has taken some distance from the intertheoretic models of reduction and now emphasizes looking at the "reduction-in-practice" in current neuroscience (Bickle 2003, 2006). He calls this approach "metascientific reductionism" to distinguish it from philosophically motivated models of reduction that are typically used in philosophy of mind.

The idea is that instead of imposing philosophical intuitions on what reduction has to be, we should examine scientific case studies to understand reduction. We should look at experimental practices of an admittedly reductionistic field, characterized as such by its practitioners and other scientists.

According to Bickle, molecular and cellular cognition – the study of the molecular and cellular basis of cognitive function – provides just the right example. The reductionist methodology of molecular and cellular cognition has two parts: (1) intervene causally into cellular or molecular pathways, (2) track statistically significant differences in the behavior of the animals (2006, 425). When this strategy is successful and a mind-to-molecules linkage has been forged, a reduction has been established. The cellular and molecular mechanisms *directly explain* the behavioural data and *set aside* intervening explanatory levels (2006, 426). Higher-level psychology is needed for describing behavior, formulating hypotheses, designing experimental setups, and so on, but according to Bickle, these are just heuristic tasks, and when cellular/molecular explanations are completed, there is nothing left for higher-level investigations to explain (2006, 428).

Metascientific reductionism does not require that the relata of reductions are formal theories, and does not lead to the problem mentioned in the end of last section. However, it is not without its share of problems, as I will show below.

4. Mechanistic explanation

The discrepancies between traditional models of reduction and actual scientific practice in psychology, neuroscience and biology have resulted in the development of alternative models. One alternative that I have just discussed is Bickle's metascientific reductionism. Another approach that has been receiving more and more attention recently is *mechanistic explanation* (*e.g.,* Bechtel & Richardson 1993, Machamer et al. 2000). In this paper I will focus on Carl Craver's (2007) recent and detailed account of mechanistic explanation.

The central claim of advocates of mechanistic explanation is that good explanations describe mechanisms (at least in neuroscience). Mechanisms are "entities and activites organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al. 2000, 3). A mechanistic explanation describes how the mechanism accounts for the *explanandum phenomenon*, the overall systemic activity (or process or function) to be explained.

For example, the propagation of action potentials is explained by describing the cellular and molecular mechanisms involving voltage-gated sodium channels, myelin sheaths, and so on. The pain withdrawal effect is explained by describing how nerves transmit the signal to the spinal chord, which in turn initiates a signal that causes muscle contraction. The metabolism of lactose in the bacterium *E. coli* is explained by describing the genetic regulatory mechanism of the *lac* operon, and so on.

5. The case of LTP

A paradigmatic example for both Bickle (2003, 43-106) and Craver (2007, 233-243) is the case of LTP (Long Term Potentiation) and memory consolidation. Both authors agree that the explanandum phenomenon is memory consolidation (the transformation of short-term memories into long-term ones), and that this is explained by describing how the relevant parts and their activities result in the overall activity - that is, by

describing the cellular and molecular mechanisms of LTP. However, the conclusions the authors draw are completely different.

According to Bickle, the case of LTP and memory consolidation is a paradigm example of an accomplished psychoneural reduction. He describes the current cellular and molecular models of LTP in detail, and argues that they are the mechanisms of memory consolidation. Furthermore, he argues that these mechanisms explain memory consolidation *directly*, setting aside psychological, cognitive-neuroscientific, *etc*., levels. This is an example of the ”intervene cellular/molecularly, track behaviorally” methodology, and in Bickle’s view a successful reduction.

What makes Bickle’s analysis ”ruthlessly” reductive is the claim that ”psychological explanations *lose their initial status as causally-mechanistically explanatory* vis-á-vis *an accomplished* (and not just anticipated) cellular/molecular explanation” (2003, 110). He argues that scientists stop evoking and developing psychological causal explanations once ”*real neurobiological* explanations are on offer”, and ”accomplished lower-level mechanistic explanations absolve us of the need in science to talk causally or investigate further at higher levels, at least in any robust ’autononomous’ sense” (2003, 111).

Craver’s analysis is quite different. He points out that the discoverers of LTP did not have reductive aspirations – they saw LTP as a component in a multilevel mechanism of memory, and after the discovery of LTP in 1973, there has been research both up and down in the hierarchy. Craver claims that the memory research program has implicitly abandoned reduction as an explanatory goal in favor of the search for multilevel mechanisms. His conclusion is that ”the LTP research program is a clear historical counterexample to those ... who present reduction as a general empirical hypothesis about trends in science” (2007, 243).

What sets Craver’s position in direct opposition to ruthless reductionism is the thesis of *causal and explanatory relevance of nonfundamental things*. That is, he argues that there is no fundamental level of explanation, and that entities of higher levels can have causal and explanatory relevance. This is in sharp contrast to Bickle’s view.

Craver's defense of the causal and explanatory relevance of nonfundamental things relies heavily on Woodward's (2003) account of causal explanation, which I will briefly present here – the details are available in Woodward's articles and books.

6. Causal explanation

A key notion for Woodward is *intervention*. An intervention can thought of as an (ideal or hypothetical) experimental manipulation carried out on some variable X (the independent variable) for the purpose of ascertaining whether changes in X are causally related to changes in some other variable Y (the dependent variable). Interventions are not only human activities, there are also "natural" interventions, and the notion of intervention can be defined with no essential reference to human agency.

Another key concept is *invariance*. Broadly speaking, a generalization or relationship is invariant if it remains intact or unchanged under at least some interventions. Suppose that there is a relationship between two variables that is represented by a functional relationship $Y = f(X)$. If the same functional relationship f holds under a range of interventions on X, then the relationship is invariant within that range. For example, the ideal gas law "$pV = nRT$" continues to hold under various interventions that change the values of the variables, and is thus invariant within this range of interventions. Invariance is a matter of degree: for example, the van der Waals force law ($[P + a/V^2][V - b] = RT$) is more invariant than the ideal gas law since it continues to hold under a wider range of interventions.

The main point is that according to Woodward, causal explanation requires appeal to *invariant generalizations*. Invariant generalizations are explanatory because they can be used to answer "what-if-things-had-been-different questions" (w-questions). For example, the ideal gas law can be used to show what the pressure of a gas would have been if the temperature would have been different. True but non-invariant generalizations like "all the coins in the pocket of Konstantin Todorov on January 25, 2008, are euros" cannot be used to answer w-questions. Only if a generalization is invariant under some range of interventions can we appeal to it to answer w-questions. In other words, causal explanatory relevance is just a matter of holding of

the right sort of pattern of counterfactual dependence between explanans and explanandum, and invariant generalizations capture these patterns.

If we accept Woodward's model of causal explanation, we see that Bickle's claims about higher-level explanations losing their status as causally/mechanically explanatory are unwarranted. In Woodward's account, things that figure in invariant generalizations have causal explanatory relevance. It is clear that in this sense nonfundamental things can have causal and explanatory relevance even when the "fundamental" cellular and molecular explanations are complete. For example, the generalizations at the higher levels of the memory consolidation mechanisms will remain invariant even after the cellular and molecular explanations are complete.

In order to counter this argument, Bickle would have to show either that the relevant higher-level generalizations are not actually invariant, or that there is something wrong with Woodward's account. The latter alternative is the more promising one. Bickle could argue that Woodward's model is simply wrong, or that there is a stronger notion of causation that applies to the cellular/molecular level. However, a notion of causation like this does not emerge from scientific evidence only (Craver's and Woodward's models are just as much based on scientific evidence as Bickle's), and Bickle seems to be reluctant to provide philosophical arguments for his views.

Furthermore, such a stronger notion of causation would inevitably lead to problems. We can always ask the question: why stop at the cellular/molecular level and not go further down to the chemical/atomic/quantum level? Bickle is conscious of this, and in fact seems to admit that it is possible that in the future causal explanations will be found at the microphysical level (2003, 156-157). This of course means that the cellular/molecular explanations are only temporarily causal explanations. It also suggests that at some point the causal explanations for all human behavior will be microphysical explanations. This kind of a notion of causal explanation strikes me as implausible and unnecessarily restrictive.

On the other hand we have Woodward's notion of causal and explanatory relevance that conforms to scientific practice and is being more and more widely accepted

among philosophers of science. The prospects of ruthless reductionism do not look very good.

7. Conclusion

In this paper, I have argued first that intertheoretic models of reduction are inappropriate for neuroscience, mainly because they focus on relations between formal theories. Then I have argued that mechanistic explanation and Woodward's theory of causal explanation taken together present a great challenge to a strongly reductionistic account of explanation in neuroscience.

Literature

Bechtel, William and Richardson, Robert C. 1993 *Discovering complexity: decomposition and localization as strategies in scientific research*, Princeton: Princeton University Press.

Bickle, John 1998 *Psychoneural Reduction: The New Wave*, Cambridge, MA: MIT Press.

Bickle, John 2003 *Philosophy and Neuroscience: A Ruthlessly Reductive Account*, Dordrecht: Kluwer Academic Publishers.

Bickle, John 2006 "Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience", *Synthese* 151, 411-434.

Churchland, Paul M. 1989 *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, Cambridge: The MIT Press.

Churchland, Patricia S. 1986 *Neurophilosophy*, Cambridge: The MIT Press.

Craver, Carl 2007 *Explaining the Brain: mechanisms and the mosaic unity of neuroscience*, Oxford: Clarendon Press.

Cummins, Robert 2000 "'How Does It Work?' vs. 'What Are the Laws?' Two Conceptions of Psychological Explanation", in: Frank Keil and Robert Wilson (eds.), *Explanation and Cognition*, Cambridge: MIT Press, 117-144.

Machamer, Peter, Darden, Lindley, and Craver, Carl 2000 "Thinking about mechanisms", *Philosophy of Science* 67, 1-25.

McCauley, Robert N. 2007 "Reduction: Models of cross-scientific relations and their implications for the psychology-neuroscience interface", in: Paul Thagard (ed.), *Handbook of the philosophy of psychology and cognitive science*, Amsterdam: Elsevier, 105-158.

Nagel, Ernest 1961 *The Structure of Science*, London: Routledge & Kegan Paul.

Richardson, Robert C. 2007 "Reduction without the structures", in: Maurice Schouten and Huib Looren de Jong (eds.), *The Matter of the Mind. Philosophical Essays on Psychology, Neuroscience and Reduction*, Oxford: Blackwell Publishing, 123-145.

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