Mechanisms, Coherence, and Theory Choice in the Cognitive Neurosciences

Stephan Hartmann[1]

Let me first state that I like Antti Revonsuo's discussion of the various methodological and interpretational problems in neuroscience. It shows how careful and methodologically reflected scientists have to proceed in this fascinating field of research. I have nothing to add here. Furthermore, I am very sympathetic towards Revonsuo's general proposal to call for a Philosophy of Neuroscience that stresses foundational issues, but also focuses on methodological and explanatory strategies.[2] In a footnote of his paper, Revonsuo complains – as many others do today – about what is sometimes called "physics imperialism". This is the view that physics dominates the philosophy of science. I am not sure if this is still the case nowadays, but it is certainly historically correct that almost all work in the field of methodology centered around cases from physics. Although this has been changing, there are still plenty of special sciences philosophers did not worry about much. Admittedly, I am myself a trained physicist and not a neuroscientist and will therefore probably be biased negatively. As it is, I will discuss some examples from physics in order to illustrate my points.

[1]I would like to thank Daniela Bailer-Jones for very helpful linguistic and substantive suggestions. A slightly extended version of this article appeared in P. Machamer et al. (eds.), *Theory and Method in the Neurosciences*. Pittsburgh: Pittsburgh University Press 2001, 70-80.

[2] It should be noted here that Patricia Churchland does, in contrast to what Revonsuo implies, indeed discuss methodological issues of neuroscience in her *Neurophilosophy* (1986). In this book (and also in subsequent publications, see Churchland and Sejnowski (1992)) she defends – following work done by W. Wimsatt (1976) - a pluralistic methodology dubbed co-evolution of theories (see p. 284f and Ch. 9). I will come back to this in Section III. It is therefore not right to claim that "neurophilosophy is regarded merely as an expression of an eliminativist-reductionist program in the philosophy of mind" because the methodological strategy of co-evolution, which Churchland defends, leaves a lot of space for different programs to develop.

My remaining comments address some of the main philosophical theses of Revonsuo's paper (especially the ones I disagree with) and is divided in three sections. The first section deals with what I think is not an acceptable genuine explanation, namely the method of visualization. The second section discusses different views of explanation in the light of neuroscience. Here I will especially focus on the presuppositions of the two dominant theories of explanation and stress, contra Revonsuo, the idea of coherence. Finally, in the third section, I address the issue of conflicting explanatory strategies for the same phenomenon in Cognitive Neuroscience and the suggested methodological consequences Revonsuo draws from this.

I.

In several parts of his paper, Revonsuo describes and praises the use of visualizations in neuroscience. Two types of visualizations deserve special attention, (1) the exposition of neural mechanisms and (2) brain imaging and mapping methods. There is no doubt that these visualizations are important tools in the actual research process; they help scientists to get a grasp of a complicated system, they are heuristically useful, represent data, and serve various didactic purposes.[3] It is a widely shared experience of scientist 'to know immediately what is going on' once one sees a good diagram. Evidently, this is why biology books are full of them.

I deny, however, that visualizations provide or facilitate explanations. Revonsuo argues for the explanatory power of visualizations when he claims - quoting Bechtel and Richardson (1993) approvingly - that idealized models "may be only partially if at all clothed in linguistic representations; instead, all kinds of visualizable diagrams and figures can often depict the component structures of, and their mutual interactions within, the biological system in question" (p...). Diagrams and figures are therefore models, and models are taken to be explanatory. I would object that, in order to make sense of a depicted mechanism, one needs all sorts of theories and theoretical models in the background. Diagrams and figures refer to, or at least hint at, the theoretical treatment in the background via various conventions shared by the users

[3] See Ruse (1990) and Wimsatt (1990).

of the depictions (Bailer-Jones 2000a). Without various theories and models and the pictorial conventions pointing to them, it is not at all clear what a diagram *means* and how it relates to the phenomenon to be explained. Besides, taking visualizations to be explanatory cannot account for the common intuition that explanations can be true or false. Precisely because visualizations employ some quite arbitrary, though convenient, conventions they elude the categories of truth and falsity. Visualizations can only be more or less useful for a certain purpose. As far as the case of brain images is concerned, Revonsuo himself pointed out how much theoretical knowledge is required to interpret the obtained pictures correctly. Moreover, measurement methods such as PET and fRMI often presuppose certain key assumptions (modularity etc.) which may not hold in nature. Taking these images literally as an explanation would, hence, be quite a dubious procedure. Let me therefore look at what scientific explanations really are.

II.

Most philosophers agree that a major aim of science is to explain phenomena. Although the concept of explanation is pretty vague we want the explanation to show (1) how the phenomenon under consideration reached its present state and (2) how it fits in a larger theoretical framework. While an acceptable answer to the first request produces *local understanding*, dealing successfully with request number two provides us with *global understanding*. Although these two requirements do not exclude each other, it remains to be seen if both can be fulfilled by the same scientific theory or model. This is not clear to start with, and philosophical theories of explanation therefore usually concentrate on one of these requirements - a task which turns out to be hard enough as the controversial debate over the last four decades or so impressively shows (Salmon 1989).

Although it is almost generally agreed that pragmatic considerations do play an enormous role in scientific explanations, Revonsuo only pays attention to the Causal/Mechanical account and the Unification account. As I will show below in section III, the neglect of pragmatic considerations is somewhat unfortunate. According to the Causal/Mechanical account (pioneered by Salmon and others, see

Salmon (1998)) a phenomenon is explained by providing a "hidden mechanism by which nature works" (p...). There are several variants of this account to be found in the literature, such as the proposals by Bechtel and Richardson, Humphreys, Salmon and Woodward.[4] Revonsuo, however, bases his reflections only on the specific account put forward by Bechtel and Richardson (1993) in a series of publications.

According to the Unification account, developed by Friedman and elaborated by Kitcher (1989), a successful explanation fits the explanandum in a coherent way in a general framework. This view, which is a distant descendent from Hempel's and Oppenheim's famous original account, supports the intuition that something is explained if it is covered by general principles. But general principles and universal laws are rare in neuroscience, which is why Revonsuo hastily concludes that unification does not play a role in this field of research. It should be noted, however, that universal principles frequently do play a role in the ordinary explanatory business of the neurosciences. This is demonstrated by the observation that one whole chapter (of seven individual contributions) in the authoritative anthology "The Cognitive Neurosciences" (Gazzaniga 1995) is devoted to "Evolutionary Perspectives". In the introduction to this chapter, the section editors Tooby and Cosmides point out that "[e]volutionary biology has a great deal to offer cognitive neuroscience. Because human and nonhuman brains are evolved systems, they are organized according to an underlying evolutionary logic. By knowing what adaptive problems a species faced during its evolutionary history, researchers can gain insight into the functional circuitry of its neural architecture" (Gazzaniga 1995, p. 1181). Besides, neuroscientific explanations ought to be consistent with all fundamental principles of physics (such as conservation laws etc.). Though these principles may not be of direct help in finding causal mechanisms, serious problems arise if a suggested mechanism violates them. Fundamental laws and principles set restrictions that may eventually even suggest a detailed 'local' explanation. Another aspect of the Unification account is even more important. Unlike the Causal/Mechanical account, the Unification account stresses the role of coherence considerations in science. I will come back to this below.

[4] The views of the three last authors are presented in Salmon and Kitcher (1989).

Antti Revonsuo reminds us that both approaches to scientific explanation make assumptions about the structure of the world that may not hold. David Lewis once asked what happens if nature is not unified. This, at first sight, seems to be a difficulty for the unificationist, and Philip Kitcher (1989) addresses this problem in detail. Revonsuo now challenges the Causal/Mechanical account by asking what would happen if the assumptions of decomposition and localization did not hold. *Decomposition* here means that the system is composed of modular subsystems, *localization* that the sub-functions within these parts can be identified (Bechtel and Richardson 1993). Indeed, the cases Revonsuo presents suggest that decomposition and localization may not be feasible in the brain.[5] Not the complete Causal/Mechanical program is at stake, however. This only means that the variant of this program Revonsuo adopted might be too narrow. Presumably, a more 'liberal' account of the Causal/Mechanical program can fully avoid these problems.

This needs to be explained. First, I do not see why it is essential to the Causal/Mechanical program that sub-functions can be localized in well-defined and spatially separated regions. In physics, parts that form functional units are often spatially separated over large distances. A typical example is superconductivity. The two correlated electrons, which form the so-called Cooper pairs (i.e. the effective degrees of freedom of a superconductor), may be localized at opposite ends of the superconductor. A "quasi-mechanical" explanation of superconductivity on the basis of the properties of Cooper pairs can nevertheless be achieved.[6] Second, Revonsuo argues that a system is decomposable if the "causal interactions within the subsystem [are] more important than those between subsystems." This does not seem to apply in physics either. According to quantum chromodynamics (QCD), the fundamental theory of strong interactions, quarks are elementary and do not have a substructure. They do, however, heavily interact inside hadrons (protons, neutrons, pions, etc.) at low and intermediate energies. In fact it is not possible to decompose the system in the laboratory and to localize individual quarks (see Hartmann (1999)). It is, however, possible to write down equations for the detailed mechanism that facilitates the strongly attractive interaction between the quarks.

[5] This makes Revonsuo's suggestion surprising that only the Causal/Mechanical program should be considered in the study of consciousness.

[6] For an interesting discussion of indeterminsitic mechanisms see Ackermann (1968; 1969).

In sum, I do not see that the Causal/Mechanical program is in trouble in neuroscience. The examples Revonsuo discusses merely suggest that the program has to be adapted better to the empirical facts. Since our theories of explanation (and especially the Causal/Mechanical account) make strong assumptions about the world, it is no surprise that we risk that some of these assumptions turn out not to hold as the scientific endeavor progresses. As far as I can see, a wider account (such as the one presented by Machamer, Darden and Craver (2000)[7]) seems to be able to deal with the problematic cases Revonsuo presents. According to Machamer, Darden and Craver a mechanism is a "regular activity of entities and their properties that are constitutive of changes from start or set up conditions to finish or termination conditions". This is not the place to flesh out this characterization in detail. It is only important to note that a mechanism in this framework simply tells us how a system evolved in time, and this seems to apply to Revonsuo's critical cases.

Having defended the Causal/Mechanical program against the charges of Revonsuo, I shall now present some of my own critical arguments. I consider all variants of the Causal/Mechanical program I know of to be incomplete because they do not stress enough the important role of *coherence* considerations in the process of establishing specific mechanisms. A proposed mechanism must cohere with our accepted background knowledge. Here it is important to note that different research programs in a preparadigmatic phase of a science (such as cognitive neuroscience) may incorporate different beliefs in their respective background knowledge. While some beliefs may be taken to be uncontroversial by all competing programs, some may be accepted by one program and dismissed by the other (such as views about the explanatory importance of the neural level or the epistemological status of folk psychology). But once a chunk of background beliefs is provisionally accepted, new beliefs in this framework should cohere with it.

Coherence is a term which is notoriously hard to define. Some plainly identify it with logical consistency. But logical consistency seems neither necessary nor sufficient for (approximate) coherence since there are usually great uncertainties in our background beliefs (especially in neuroscience, as Revonsuo stresses in his paper).

[7] See also Craver´s contribution to this volume.

So, if some propositions of a belief system are uncertain, a contradiction resulting from integrating a new proposition in the system would not destroy the coherence of the whole system. Besides, logical consistency does not seem to be a very good guide to point to a desired mechanism. Too many mechanisms do the job to bring the system from here to there, but not all of them are accepted – for good reasons.[8]

Although a new mechanism can suggest a radically new aspect of our world, it will still be linked to other parts of our belief system. The question therefore matters how well this new mechanism coheres with the rest of this system. In order to make this claim precise, requires getting in the deep epistemological waters of defining what coherence means. Here is what Lawrence BonJour has to say about this:

What then is coherence? Intuitively, coherence is a matter of how well a body of belief 'hangs together': how well its component beliefs fit together, agree or dovetail with each other, so as to produce an organized, tightly structured system of beliefs, rather than either a helter-skelter collection or a set of conflicting subsystems. It is reasonably clear that this 'hanging together' depends on the various sorts of inferential, evidential, and explanatory relations which obtain among the various members of a system of belief, and especially on the more holistic and systematic of these (BonJour 1985, p. 93).

Following this line of thought we would then accept a proposed mechanism if it makes a given system of beliefs more coherent or at least if it does not make it less coherent. This seems to be intuitively clear and Revonsuo himself discusses a couple of examples where coherence considerations play a role. It is obvious, for example, that the models of neural systems at several levels of description must cohere. Revonsuo here mentions "synapses and synaptic transmission, single neuron morphology and electrophysiology, neural connectivity and organization in sensory systems and the central nervous system, cytoarchitectonics of the cerebral cortex, macroanatomy of the brain, and so forth" (p...). It would indeed be a miracle if models of so many interrelated levels fitted together without using coherence as an important constraint in theory construction. Another example for the role of coherence

[8] Bailer-Jones (2000b) develops the role of causal mechanisms in a similar direction, but perhaps misleadingly talks about consistency rather than coherence.

considerations in the neurosciences is the already mentioned interpretation of pictures obtained by PET or fMRI measurements. Acceptable results of such measurements should cohere with the other assumptions of the "experimental paradigm." Since "any one of [these assumptions] might be wrong" (p...), this turns out to be a difficult task.

Although all this seems to be intuitively clear and scientists use such a principle in their daily work, it is nevertheless a big problem in formal philosophy to provide a quantitative measure (say a number between zero and one) for the coherence of a belief system. I suggest that this is best done in a probabilistic framework. Luc Bovens and myself have shown elsewhere that such a qualitative measure of coherence can be obtained if one specifies coherence to be a confidence boosting property of a set of beliefs and uses the mathematical theory of Bayesian Networks which is well-known in artificial intelligence research. I cannot go into details here and refer the reader to the literature (Bovens and Hartmann (forthcoming)).

I propose that the search for coherence is an important guiding principle in finding acceptable explanations in neuroscience.[9] Most of these explanations might indeed be causal/mechanical (although I doubt that all are), but if a suggested mechanism does not fit in a set of background assumptions, it will have a hard time to be accepted by the scientific community. Maybe there is nothing more to gain from science than a picture of the world which is as coherent as possible. Even if the Causal/Mechanical account really has to be given up at some point, the idea of coherence will still play a dominant role in scientific theorizing.

I therefore doubt that there are really two alternative views of explanation which may complement each other and even coexist in science. Recall Salmon's (1998, p. 73f) story of the friendly physicist. Salmon states that there are two equally acceptable explanations for the phenomenon that a balloon moves forward in an airplane when it accelerates. The causal/mechanical explanation goes with an intuitive story about the movement of the gas molecule, while the explanation-as-unification applies Einstein's equivalence principle. Again I would like to stress that local laws such as the ones

[9] The relation between the notions of coherence and explanation is also discussed in Bartelborth (1999).

which govern the behavior of gas particles would not be accepted as explanatory if they contradicted general principles such as the equivalence principle and if they were not part of a coherent larger framework. It is the interaction between these two approaches (a bottom-up approach and a top-down approach) which seems to characterize science and its ability to explain best. Frequently, general considerations help us to obtain 'local ' knowledge, and the analysis of specific mechanisms may provide 'global' knowledge.

III.

According to Revonsuo, there is a deep confusion in the foundations of the new discipline cognitive neuroscience because conflicting explanatory strategies crash at the interface of the old disciplines cognitive science and neuroscience. While neuroscience aims at explaining mental or cognitive levels in a mechanistic way, cognitive science is committed to functionalism and the autonomy of psychology. In Revonsuo's view functionalism implies "the total independence of psychological and neural levels of description." This is certainly too strong a claim since even if there are multiple ways to realize an algorithm, it must still be shown or made plausible that the human brain is able to do so.[10]

Even if we accept, however, for the sake of argument that there is a conflict (after all, eliminative-materialists want to get rid of folk psychology at the end of the day!), the question can be asked what to do in this situation. Revonsuo has a radical proposal: He suggests to abandon the cognitivist-functionalist strategy and argues for a "full-scale" mechanistic program to be applied to cognitive neuroscience.[11]

Again I consider this to be too radical a proposal and I am not sure if Revonsuo is serious about what he writes. First, Revonsuo himself pointed out that the Causal/Mechanical program might be in trouble. Second, it is not clear to me that the

[10] Another argument against the multiple-realizations argument is given in Bechtel (1999).

[11] Revonsuo advises this strategy even for the case of consciousnes, although he only formulates vaguely that it is his belief that the Causal/Mechanical account is superior.

evidence for the Causal/Mechanical account is so overwhelming and that the evidence for the cognitive science account is so poor that one definitely has to abandon the latter and follow the first. Third, and most importantly, research programs which evolve in parallel can be of great advantage for each other. Churchland explains why:

[T]heories at distinct levels often co-evolve [...], as each informs or corrects the other, and if a theory at one stage of its history cannot reduce a likely candidate at a higher level, it may grow and mature so that eventually it does succeed in the reductive goal. In the meantime the discoveries and problems of each theory may suggest modifications, developments, and experiments for the other, and thus the two evolve towards a reductive consummation. (Churchland 1986, p. 384)

This quotation shows that even a strong proponent of an eliminativist-materialistic ontology can live with methodological pluralism (and advice it). In any case, Revonsuo's thesis seems to me to be by far too strong.

One may wonder if the pluralism I just defended is a problem for the aim formulated in the last section, to reach for coherent theories. Obviously, if we include cognitive science and neuroscience in our theory of the world, the resulting system will not be considerably coherent. Having a measure of coherence might suggest, pace Revonsuo, to only follow the program that seems to be the more coherent of the two. In accordance with the idea of co-evolution I would, however, not advice this. We have to live with the fact that some sciences are still in a pre-paradigmatic stage and in this situation it is best to let both programs grow, profiting both from cross-fertilization etc. Coherence now should only play a role within a given framework (say, cognitive science, plus background knowledge from other sciences, plus experimental data etc.). The competing research programs should not be taken into account here. Inconsistencies between different research programs do not matter at this stage of theory development. This, in a way, is a pragmatic component that enters here. Once we have chosen a certain framework we can reach for coherent theories. The question of when a research program has to be given up needs, of course, further investigation.

What is, finally, the exact place of psychology? This remains an open question until we have a satisfactory account of our cognitive functions. Until then, there is nothing wrong with a pluralism of different explanations for the same phenomenon.[12] It is legitimate to approach a given subject matter from different directions without worrying too much about their mutual consistency. In this same spirit Patricia Churchland once advised us to follow the research principle: "Let a thousand flowers bloom", but before I exit the scientific arena entirely, I had better stop.

References

Ackermann, R. (1968), "Mechanism and the Philosophy of Biology," *Southern Journal of Philosophy* 6: 143-151.

Ackermann, R. (1969), "Mechanism, Methodology, and Biological Theory," *Synthese* 20: 219-229.

Bailer-Jones, D. M. (2000a, to appear), "Sketches as Mental Reifications of Theoretical Scientific Treatment," in: M. Anderson, B. Meyer and P. Olivier, *Diagrammatic Representation and Reasoning*. London: Springer.

Bailer-Jones, D. M. (2000b, to appear), "Modelling Extended Extragalactic Radio Sources," *Studies in History and Philosophy of Modern Physics* 31B.

Bartelborth, T. (1999), "Explanatory Coherence," *Erkenntnis* 50: 209-224.

Bechtel, W. and R. Richardson (1993), *Discovering Complexity*. Princeton: Princeton University Press.

Bechtel, W. and J. Mundale (1999), "Multiple Realizability Revisited: Linking Cognitive and Neural States," *Philosophy of Science* 66: 175-207.

[12] For a defense of this view for biology see Mayr (1997).

BonJour, L. (1985), *The Structure of Empirical Knowledge*. Cambridge, Mass.: Harvard University Press.

Bovens, L. and S. Hartmann (forthcoming), “The Riddle of Coherence,” preprint, Boulder and Konstanz.

Churchland, P.S. (1986), *Neurophilosophy*. Cambridge, Mass.: MIT Press.

Churchland, P.S. and T. Sejnowski (1992), *The Computational Brain*. Cambridge, Mass.: MIT Press.

Hartmann, S. (1999) “Models and Stories in Hadron Physics”, in: M. Morrison and M. Morgan (eds.), *Models as Mediators. Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press, pp. 326-346.

Gazzaniga, M. (1995), *The Cognitive Neurosciences*. Cambridge, Mass.: MIT Press.
Kitcher, P. (1989) “Explanatory Unification and the Causal Structure of the World,” in: Kitcher and Salmon (1998), pp. 410-505.

Kitcher, P. and W. Salmon (1989), *Scientific Explanation*. Minneapolis: University of Minnesota Press.

Machamer, P., L. Darden and C. Craver (2000, to appear), “Thinking about Mechanisms,” *Philosophy of Science* 67.

Mayr, E. (1997), *This is Biology. The Science of the Living World*. Cambridge, Mass.: Harvard University Press.

Ruse, M. (1991), “Are Pictures Really Necessary? The Case of Sevell Wright’s ’Adaptive Landscapes’,” in: A. Fine, M. Forbes, and L. Wessels (eds.), *PSA 1990, Vol. 2*, East Lansing: The Philosophy of Science Association, pp. 63-77.

Salmon, W. (1989), *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press. (This text is also contained in Kitcher and Salmon (1989), pp. 3-219.)

Salmon, W. (1998), *Causality and Explanation*. Oxford: Oxford University Press.

Wimsatt, W. (1976) “Reduction, Levels of Organization, and the Mind-Body Problem,” in: G. Globus, G. Maxwell, and I Savodnik (eds.), *Consciousness and the Brain*. New York: Plenum Press, pp. 199-267.

Wimsatt, W. (1991), “Taming the Dimensions. Visualizations in Science,” in: A. Fine, M. Forbes, and L. Wessels (eds.), *PSA 1990, Vol. 2*, East Lansing: The Philosophy of Science Association, pp. 111-135.