On The Scope of Mechanistic Explanation in Cognitive Sciences

Anna-Mari Rusanen (anna-mari.rusanen@helsinki.fi)
Department of Philosophy, History, Art and Culture Studies,
PO BOX 24
00014 University of Helsinki FINLAND

Otto Lappi (otto.lappi@helsinki.fi)
Institute of Behavioural Sciences,
PO BOX 9
00014, University of Helsinki FINLAND

Abstract

Computational explanations focus on information processing tasks of specific cognitive capacities. In this paper, we argue that there are at least two different kinds of computational explanations; the interlevel and the intralevel ones. Moreover, it will be argued that neither interlevel nor intralevel computational explanations can be subsumed under the banner of standard mechanistic explanations. In the case of interlevel explanations, the problem is the direction of explanation, and in the case of intralevel explanations, the problem are the dependencies that the explanations track. Finally, it is argued that in the context of explanation of cognitive phenomena, it may be necessary to defend more liberal and pluralistic views of explanation, which would allow that there are also some non-mechanistic forms of explanation.

Keywords: computational explanation; mechanistic explanation; computation; Marr

Introduction

Computational explanations focus on information processing required in exhibiting specific cognitive capacities, such as perception, reasoning or decision making. At an abstract level, these computational tasks can be specified as mappings from one kind of information to another.

These explanations can increase our understanding of a cognitive process at least in three ways: (i) they can explain a certain cognitive phenomenon in terms of fundamental rational or mathematical principles governing the information processing task faced by a system, or (ii) they can explain by describing the formal dependencies between certain kinds of tasks and certain kinds of information processing requirements. Moreover, in many computational accounts[1] it is often assumed that (iii) computational explanations can explain the phenomenon in terms of its implementation in more primitive constituent processes.

In recent years, a number of philosophers have proposed that computational explanations of cognitive phenomena could be seen as instances of mechanistic explanation (Piccinini 2004; 2006b; Sun 2008; Kaplan, 2011; Piccinini & Craver 2011).

In what follows, we will argue that while fulfilling these epistemic needs is essential in computational explanation in the cognitive sciences, only the last mode of explanation conform to the mechanists' way of thinking what genuine mechanistic explanation is.

Thus, we conclude that either philosophers of cognitive science need to embrace non–mechanistic computational explanations, or extend the scope of what counts as "mechanistic" explanation in cognitive science.

Computational Explanations and Mechanistic Explanation

Within the last ten years, a growing number of philosophers have defended the view that computational explanations are mechanistic explanations (Piccinini 2004; Kaplan 2011; Craver & Piccinini 2011). For example, according to Piccinini (2004, 2006a, 2006b) computing mechanisms can be analyzed in terms of their component parts, their functions, and their organization. For Piccinini, a computational explanation is then "a mechanistic explanation that characterizes the inputs, outputs, and sometimes internal states of a mechanism as strings of symbols, and it provides a rule, defined over the inputs (and possibly the internal states), for generating the outputs" (Piccini 2006b).

According to this mechanistic account, the goal of computational explanation is to characterize the functions that are being computed (the what) and specify the algorithms by which the system computes the function (the how). In other words, the idea is that an information processing phenomenon would be explained by giving a sufficiently accurate model of how hierarchical causal systems composed of component parts and their properties sustain or produce the phenomenon[2].

[1] For instance, Piccinini 2006a,2006b, Kaplan 2011. See also Shagrir 2010 for discussion.

[2] Constructing an explanatory mechanistic model thus involves mapping elements of a mechanistic model to the system of interest, so that the elements of the model correspond to identifiable constituent parts with the appropriate organization and causal powers to sustain that organization. These explanatory models should specify the initial and termination conditions for the mechanism, how it behaves under various kinds of interventions, how it is integrated with its environment, and so on.

This kind of mechanistic "computational" explanations track causal dependencies at the level of cognitive performances. They correspond to explanations which David Marr (1982) called "algorithmic" explanations. However, as we have argued earlier (Rusanen & Lappi 2007; Lappi & Rusanen 2011), it is not obvious, whether this mechanistic account can be extended to cover computational explanations in Marr´s sense[3].

In Marr´s trichotomy, computational explanations specifies what are the information processing tasks, and what is computed and why. Computational explanations give an account of the tasks that the neurocognitive system performs, or problems that the cognitive system in question is thought to have the capacity to solve, as well as the information requirements of the tasks (Marr, 1982).

This level of explanation is also the level, whereby the appropriateness and adequacy (for the task) of mappings from representations to others are assessed (cf. Marr, 1982). For example, in the case of human vision, one such task might be to faithfully construct 3D descriptions of the environment from two 2D projections. The task is specified by giving the abstract set of rules that tells us what the system does and when it performs a computation. This abstract computational theory characterizes the tasks as mappings, functions from one kind of information to another. It constitutes, in other words, a theory of competence for a specific cognitive capacity - vision, language, decision making etc.

The Interlevel and The Intralevel Computational Explanations

It is important to distinguish two different types of computational explanations. Firstly, there are *interlevel computational explanations*, which explain by describing, how the possible behavior or processes of a system is governed by certain information processing principles, rather than explain how certain algorithms compute certain functions. These computational explanations display the function that the mechanism computes and they explain and why this function is appropriate for a given cognitive task.

Some of our critics, such as Milkowski, have claimed that we see these interlevel computational explanations as "systemic explanations that show how a cognitive system can have some capacities" (Milkowski 2013, p. 107). However, we do not defend such a position. We do not claim that computational explanations explain *how* a cognitive system *can have some capacities*. Instead, what we claim is that interlevel computational explanations explain *why* and *how* certain *principles govern* the possible behavior or processes of the system.

In this sense, interlevel explanations explain the behavior of mechanisms at the algorithmic and implementation levels. In such explanations, the explanans is at the "upper" computational level, and the explananda are at the "lower" algorithmic or performance levels. For example, if one considers, *why* certain synaptic change is such-and-such, answers are often something like "*because it serves to store the value of x needed in order to compute y*. Or, *why* is the wiring in this ganglion such-and-such? Because it computes, or approximates computation of x. In other words, phenomena at the lower levels are explained by their appropriateness of the mechanism for the computational tasks.

Secondly, there are computational explanations, which are rather *intralevel* than interlevel explanations. In short, these explanations track formal dependencies between certain kinds of information processing tasks, and they explain by describing certain kinds of information processing requirements at the level of cognitive competences.

There are different views about the nature of the formal dependencies, which are tracked by these computational explanations. Some take it that the dependencies can be described intentionally i.e. in terms of informational content, while some other, such as Egan (1992) argues that computational explanations track appropriate mathematical dependencies by specifying the mathematical input-output-functions that is being computed. There are also some pluralistic views; for instance Shagrir (2010) defends the view that there are actually two different types of formal dependencies; the "inner" and the "outer" ones. According to Shagrir (2010) the inner formal dependencies are formal relations between inputs and outputs, and the outer formal dependencies are mathematical relations between "what is being represented by the inputs and outputs". These formal dependencies are abstracted from representational contents, which correspond for example certain features of physical environment.

So, there are at least two different kinds of computational explanations; the interlevel and the intralevel ones. In the following sections, we will argue that neither interlevel nor intralevel computational explanations can be subsumed under the banner of standard mechanistic explanations. In the case of interlevel explanations, the problem is the direction of explanation (Rusanen & Lappi 2007), and in the case of intralevel explanations, the problem are the dependencies that the explanations track (Rusanen 2014).

Inter-level Computational Explanations: The Problem of Direction

In a nutshell, the problem for standard mechanistic accounts of interlevel explanations goes as follows: In standard accounts (constitutive) mechanistic explanations are characterized in such a way that in inter-level computational explanations, the *explanans is at a lower level than the explanandum.* For example Craver (2001, p. 70, emphasis

[3] Although Marr´s notion of computational explanation is sometimes thought to be "outdated" and "oldfashioned", it still plays an important role in cognitive and cognitive neurosciences. For example, there is interesting work being done in theoretical neuroscience and cognitive modeling within this framework in the domains of vision, language, and the probabilistic approach to cognition (for overviews, see Anderson 1991; Chater 1996; Chater et al. 2006).

added) notes that " (Constitutive) explanations are *inward* and *downward* looking, looking within the boundaries of X to determine the *lower* level mechanisms by which it can Φ. The explanandum… is the Φ-ing of an X, and the explanans is a description of the organized σ-ing (activities) of Ps (still lower level mechanisms)."

In those explanations, phenomena at a higher level of hierarchical mechanistic organization are explained by their lower-level constitutive causal mechanisms but not vice versa (Craver 2001, 2006; Machamer & al, 2000). For example, under this interpretation a cognitive capacity would be explained by describing implementing mechanisms at algorithmic or implementing level. But, in inter-level computational explanations, the competence explains performance i.e. explanans is at the level of cognitive competences, and the explanandum is at the level of performances. In other words, these inter-level computational explanations proceed *top-down*, while constitutive mechanistic explanations are typically characterized in such a way that they seem always to be bottom-up explanations. Thus, computational explanations are not constitutive mechanistic explanations in the standard sense.

One might argue that this analysis ignores the possibility that computational explanations are *contextual* rather than *constitutive* mechanistic explanations. In the mechanistic terminology, the contextual explanations explain how the "higher-level" mechanism constrains what a lower level mechanism does, and one computational mechanism can be a component of a larger computational system, while the latter serves as the contextual level for the former. For example Bechtel seems to accept this position, when he remarks that "since (marrian) computational explanations address what mechanisms are doing they focus on mechanisms "in context"" (Bechtel 2008, p. 26).

Now, *if* computational explanations was contextual explanations, *then* our argument would fail. Namely, if computational-level explanations were contextual explanations, and if contextual explanation is a subspecies of standard mechanistic explanations, then computational level explanations would be a subspecies of mechanistic explanations.

However, it is possible to argue that computational explanations are not contextual explanations in the standard mechanistic sense. For instance, Craver characterizes contextual explanations as explanations, which "refer to components *outside* of X" and are "upward looking because they contextualize X within a higher level mechanism". On this view, a description of how a cognitive system "behaves" in its environment, or how an organization of a system constraints the behavior of its components, require a spatiotemporal interpretation for the mechanisms. But, as we argued in 2011, computational explanations do not necessarily refer to *spatiotemporally implemented* higher-level mechanisms, and they do not involve spatiotemporally implemented components "outside of (spatiotemporally implemented) X". Instead, they refer to abstract "mechanisms", which are not causally or spatiotemporally implemented.

In other words, the problem is that in standard mechanistic accounts, in contextual explanations the "contexts" are expressed in causal and spatiotemporal terms, *not* in terms of information processing at the level of computational competences. Crucially, this kind of view conceives contextual explanations as a kind of systemic explanations, in which the uppermost level of the larger mechanism will still remain non-computational in character.

For this reason, computational explanations are not these "systemic" contextual explanations. In contrast, we claim, computational explanations involve abstract mechanisms, which are not causally, but logically governing the behavior of the mechanisms at the lower levels.

Intra–level Computational Explanations: The Problem of Dependencies

Now, let´s move to the intralevel computational explanations. Why cannot they be seen as standard mechanistic explanations? Well, the answer is that they simply track different kinds of dependencies. While algorithmic and implementation level explanation track causal or constitutive dependencies at the level of cognitive or neural performances, intra-level computational explanations track formal dependencies between certain kinds of information processing tasks at the level of cognitive competences.

Because of this, these different modes of explanation are not necessarily logically dependent on each other. Thus the computational explanations at the highest level may be formulated independently of assumptions about the algorithmic or neural mechanisms which perform the computation.

Some of our critics, such as Kaplan (2011) and Piccinini (2009) remark that our position can be seen as a typical example of "computational chauvinism", according to which computational explanations of human cognitive capacities can be constructed and confirmed independently of details of their implementation in the brain.

Indeed, we defend the view that computational explanations can be in principle - if not in practice - constructed largely autonomously with respect to the algorithmic or implementation levels below. That is: *computational problems* of the highest level may be formulated independently of assumptions about the algorithmic or neural mechanisms which perform the computation (Marr 1982; see also Shapiro 1997; Shagrir 2001). Because the performance and competence- level computational explanations track different kinds of dependencies, these different modes of explanation are not necessarily logically dependent on each other. Hence, if this is computational chauvinism, then we are computational chauvinists.

However, Kaplan (2011) claims that while we highlight the independence of computational explanations, we forget something important Marr himself emphasized. Namely,

Kaplan remarks that even if Marr emphasized that the same computation might be performed by any number of algorithms and implemented in any number of diverse hardwares, Marr´s position changes when he "addresses the key explanatory question of whether a given computational model or algorithmic description is appropriate for the specific target system under investigation" (Kaplan 2011, p.343).

Is this, really, an argument against our position? As Kaplan himself remarks, Marr rejects "the idea that any computationally adequate algorithm (i.e., one that produces the same input-output transformation or computes the same function) is equally good as an explanation of how the computation is performed in that particular system" (Kaplan 2011 p.343).

But then, we are not talking about competence level explanations anymore. When the issue is how the computation is performed in the particular system, such as in human brains, then the explanation is given in terms of algorithmic or neural processes, or mechanisms, if you will. Then, *naturally*, the crucial issue is what kinds of algorithms are possible for a certain kind of system, or whether the system has structural components that can sustain the information processing that the computational model posits at the neural level. If one aims to explain how our brains are able to perform some computations, then – of course – one should take the actual neural implementation and the constraints of the possible neurocognitive architecture into account as well.

But given this, these kinds of explanations are explanations at the algorithmic or performance level, not at the computational or competence level. Because of this, we also find position defended by Piccinini & Craver (2011) problematic. Piccinini and Craver (ibid) argue that in so far computational explanations do not describe how the computational system "actually works" i.e. describe "how the information is encoded and manipulated" in implementing system, they are mere how possibly-explanations. In our understanding, this depends on the explanatory questions. If, for example, the aim is to explain, *how* certain kind of information processing task is actually solved in human brains, and if the explanations does not describe how it actually happens, it is a how possibly-explanation. But, it is a how possibly explanation *at the performance level, not at the competence level.*

For this reason, the remark that computational explanations do not describe how the computational system "actually works" is not an argument against the logical independence of the computational level explanations.

The Explanatory Status of Computational Explanations

A more problematic issue is to what extent computational explanations are explanatory after all. Although Milkowski may partially misinterpret our position, he still raises an important question concerning the explanatory character of computational explanations (Milkowski 2012, 2013).

If computational explanations are characterized as explanations which answer questions such as: "What is the goal of this computation?", it may be claimed that we fail to make a distinction between task analysis and genuine explanations.

A task analysis breaks a capacity of a system into a set of sub-capacities and specifies how the sub-capacities are (or may be) organized to yield the capacity to be explained. Obviously, if computational explanations are mere descriptions of computational tasks, then they are not explanations at all.

However, computational explanations are clearly more than mere descriptions of computational tasks, because they describe formal dependencies between certain kinds of tasks and certain kinds of information processing requirements. If these formal dependencies are such that descriptions of them not only offer the ability to say how the computational layout of the system actually is, but also the ability to say how it would be under a variety of circumstances or interventions, they can be counted as explanatory[4].

In other words, if these descriptions answer questions such as "Why does this kind of task create this kind of constraint rather than that kind of constraint?" by tracking such formal dependencies which can explain what makes the difference, then these descriptions can be explanatory.

Obviously, computational explanations of this sort are not causal explanations. However, in the context of explanation of cognitive phenomena, it may be necessary to defend more liberal and pluralistic views of explanation, which would allow that there are also some non-causal forms of explanation.

We agree with mechanists that when we are explaining how cognitive processing actually happens for example in human brains, it is a matter of causal explanation to tell how the neuronal structures sustain or produce the information processing in question. However, we still defend the view that there are other modes of explanation in cognitive sciences as well.

Discussion: The Scope of Mechanistic Explanation

Some explanations of cognitive phenomena can be subsumed under the banner of "mechanistic explanation". Typically those explanations are neurocognitive explanations of how certain neurocognitive mechanisms produce or sustain certain cognitive phenomena, but also some psychological explanations can be seen as instances of mechanistic explanations. Moreover, if a more liberal interpretation for the term mechanism is allowed, then *some* computational or competence level explanations may also qualify as mechanistic explanations (Rusanen & Lappi 2007; Lappi & Rusanen 2011).

[4] This is a non-causal modification of the Woodward´s manipulationist account of explanation (Woodward 2003). For a similar treatment of Woodward, see Weiskopf 2011.

Nevertheless, we think that there are compelling reasons to doubt whether mechanistic explanation can be extended to cover *all* cognitive explanations. There are several reasons for this plea for explanatory pluralism: Firstly, it is not clear whether all cognitive systems or cognitive phenomena can be captured mechanistically. Mechanistic explanations require that the system can be decomposed i.e. analyzed into a set of possible component operations that would be sufficient to produce or sustain the phenomenon in question (Bechtel & Richardson 1993). Typically a mechanism built in such a manner will work in a sequential order, so that the contributions of each component can be examined separately (Bechtel & Richardson 1993).

However, in cognitive sciences there are examples of systems – such as certain neural nets – which are not organized in such a manner. As Bechtel and colleagues remark, the behavior of these kinds of systems cannot be explained by decomposing the systems into subsystems, because the parts of the networks do not perform any activities individually that could be characterized in terms of what the whole network does (Bechtel & Richardson 1993; Bechtel 2011, 2012). Hence, it is an open question to what extent the behavior of these kinds of systems can be explained mechanistically. At the very least, it will require adopting a framework of mechanistic explanation different from the one that assumes sequential operation of decomposable parts (Bechtel 2011, 2012; Bechtel & Abrahamsen 2011).

Moreover, Von Eckardt and Poland (2004) raise the question to what extent the mechanistic account is appropriate for those explanations which involve appeal to mental representations or to the normative features of certain psychopathological phenomena. Although we find Von Eckardt and Poland´s argumentation slightly misguided, we still think that it is important to consider the normative aspects of cognitive phenomena. Cognitive systems are, after all, adaptive systems which have a tendency to seek "optimal", "rational" or "best possible" solutions to the information processing problems that they face. Because of this, cognitive processes are not only goal-directed, but also normative. It is not clear how well this normative aspect of cognitive systems can be captured by mechanistic explanations.

Thirdly, some philosophers have paid attention to the fact that there are examples of explanatory computational models in cognitive sciences which focus on the flow of information through a system rather than the mechanisms that underlie the information prosessing (Shagrir 2006, 2010). Along similar lines, Weiskopf (2011) argues that there is a set of "functional" models of psychological capacities which are both explanatory and non-mechanistic.

Finally, in recent years cognitive scientists have raised the possibility that there are some universal, law-like principles of cognition, such as the "principle of simplicity", "universal law of generalization" or the "principle of scale-variance" (Chater & al 2006; Chater & Vitanyi 2003). Chater and colleagues (ibid.) argue that it is possible to explain many cognitive phenomena, such as certain forms of linguistic patterns, or certain types of inductive generalizations, by combining these principles.

These explanations are "principle based" rather than mechanistic explanations. Moreover, Chater and colleagues seem to suggest the mechanistic models of these phenomena may actually be derived from these general principles, and explanations that appeal to these general principles provide "deeper" explanations than the mechanistic explanations (Chater & Brown 2008). It is possible, that many of the so called computational level explanations turn out to be instances of these principle-based explanations rather than instances of mechanistic explanations.

In sum, taken together these diverse claims seem to imply that there is not a single, unified mode of explanation in cognitive sciences. Instead, they seem to suggest that cognitive sciences are examples of those sciences which utilize several different modes of explanation, only some of which can be subsumed under the mechanistic account of explanation.

Obviously, mechanistic explanation is a powerful framework for explaining the behavior of complex systems, and it has demonstrated its usefulness in many scientific domains. Also, many successful theories and explanations in cognitive sciences are due to this mechanistic approach. However, this does not imply that it would be the only way to explain complex cognitive phenomena.

Concluding Remarks

In this paper, we have argued that there are at least two different kinds of computational explanations; the interlevel and the intralevel ones. Moreover, we have argued that neither interlevel nor intralevel computational explanations can be subsumed under the banner of standard mechanistic explanations. In the case of interlevel explanations, the problem is the direction of explanation, and in the case of intralevel explanations, the problem are the dependencies that the explanations track.

Obviously, computational explanations of this sort are not causal explanations. However, in the context of explanation of cognitive phenomena, it may be necessary to defend more liberal and pluralistic views of explanation, which would allow that there are also some non-causal forms of explanation.

References

Anderson, J. 1991b. Is Human Cognition Adaptive? *Behavioral and Brain Sciences*, 14: 471- 457.

Bechtel, W. 2008. Mental mechanisms: Philosophical perspectives on cognitive neuroscience. London: Routledge University Press.

Chater, N. 1996. Reconciling Simplicity and Likelihood Principles in Perceptual Organization. *Psychological Review*, 103(3): 566-581.

Chater, N., & Vitanyi, P. M. B. 2003. The generalized universal law of generalization. *Journal of Mathematical Psychology*, 47: 346-369.

Chater, N., Tenenbaum, J. B., & Yuille, A. 2006. Probabilistic Models of Cognition: Conceptual Foundations. *Trends in Cognitive Sciences*, 10, 287-291.

Chater, N. & Brown, G. From Universal Laws of Cognition to Specific Cognitive Models. Cognitive Science, 32, 36-67.

Craver, C.F. 2001. Role functions, Mechanisms and Hierarchy. Philosophy of Science, 68, 53-74.

Craver, C.F. 2006. When Mechanistic Models Explain. Synthese, 153: 355-376.

Kaplan, D. 2011. Explanation and description in computational neuroscience. Synthese, 183 (3): 339-373

Lappi, O. & Rusanen, A-M. 2011. Turing Machines and Causal Mechanisms in Cognitive Sciences, In P. McKay Illari, F. Russo & J. Williamson, (eds.) Causality in the Sciences. Oxford: Oxford University Press. 2011: 224-239

Machamer, P. K., Darden, L., & Craver, C. 2000. Thinking About Mechanisms. Philosophy of Science, 67: 1-25.

Marr, D. 1982. Vision: A Computational Investigation into the Human Representation of Visual Information. San Francisco: W.H. Freeman.

Milkowski, M. 2012. Limits of Computational Explanation of Cognition. In Müller, V. (ed.) Philosophy and Theory of Artificial Intelligence. Springer.

Milkowski, M. 2013.Explaining the Computational Mind. MIT Press.

Piccinini, G. 2004a. Functionalism, Computationalism and Mental Contents. Canadian Journal of Philosophy, 34, 375-410.

Piccinini, G. 2006a. Computational Explanation and Mechanistic Explanation of Mind. In M. DeCaro, F. Ferretti & M. Marraffa (Eds.), Cartographies of the Mind: The Interface Between Philosophy and Cognitive Science. Dordrecht: Kluwer.

Piccinini, G. 2006b. Computational Explanation in Neuroscience. Synthese, 153, 343-353.

Piccinini, G. & Craver, C. 2011. Integrating psychology and neuroscience: functional analyses as mechanism sketches. Synthese, 183 (3):283-311.

Piccinini, G. 2011. Computationalism, in *Oxford Handbook of Philosophy of Cognitive Science*, Eric Margolis, Richard Samuels, and Stephen Stich, eds., Oxford: Oxford University Press 2011: 222-249.

Rusanen, A-M. & Lappi, O. (2007). The Limits of Mechanistic Explanation in Neurocognitive Sciences. In Vosniadou, S., Kayser, D. & A. Protopapas, (Eds) *Proceedings of the European Cognitive Science Conference 2007*. Howe: Lawrence Erlbaum Associates. 2007: 284-289.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Shagrir, O. 2010. Brains as Analog-Model Computers. *Studies in the History and Philosophy of Science*, 41(3): 271-279.

Shapiro, L. 1997. A Clearer Vision. *Philosophy of Science*, 64, 131-153.

Weiskopf, D. 2011. Models and mechanisms in psychological explanation, *Synthese*, 183: 313-38.

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