Plausibility versus richness in mechanistic models[1]

Introduction

Over the past decade, philosophers have increasingly emphasized the role mechanistic explanations play in the life sciences (Machamer et al. 2000; Glennan 2002; Tabery 2004; Bechtel and Abrahamsen 2005). Though the notion MECHANISM is inherited from early modern philosophy, its meaning has departed from the traditional understanding of mechanisms: rather than physical systems involving gears, flywheels, springs or colliding particles, they are viewed as conceptual collections of entities and activities that are organized such as to produce a certain phenomenon.[1] Researchers are thought to *explain* this phenomenon by describing the responsible mechanism. In other words, explaining consists of providing a model of the mechanism. Examples of disciplines in the life sciences where this explanatory strategy is used include among others neuroscience (Craver 2007; Bechtel 2008) and biology (Darden & Craver 2002).

Although I agree that researchers in the life sciences typically explain a phenomenon by providing a model of the responsible mechanism, I think that the recent debate on mechanistic explanations conflates two features of models that, for reasons I will establish below, are better kept distinct: their *plausibility* and their *richness of information*. By plausibility, I mean the degree of probability that a model is accurate in the existence of, and distinctions between, the various entities and activities it postulates,[2] while richness concerns the degree of detail a model provides in its description of a mechanism's entities and activities. The conflation of these two features in the debate on mechanistic explanations is undesirable, as it has led some of the participants to view both of them as necessary for a model to have explanatory power, while in fact only one, namely plausibility, is required. Richness, although a virtue for many other reasons, is not necessary for a model to be explanatory: there are models that say next to nothing about a mechanism's entities but still have considerable explanatory power. To put it plainly, the conflation of plausibility and richness leads one to discard as non-explanatory models that quite clearly are.

This paper is divided into three sections. In section one, I will show how the debate on mechanistic explanation confuses plausibility with richness, by considering how mechanistic models relate to the more traditional functional models of the 1970s and -80s. As a focal point, I will use Craver's continuum from how-possibly to how-actually models (Craver 2006)[3], although the

[1] The research for this paper was supported by the Research Fund Flanders (FWO) through project nr. G.0031.09.

confusion applies more generally to the debate on mechanistic explanations. In section two, I will briefly make the case that plausibility and richness can vary independently of each other in more than one way, so that there is at least a conceptual reason to keep them apart. Finally, the third section brings out the true cost of the confusion, by providing an example of a model that is explanatory, but offers no details about the entities of the mechanism it is describing. This counterexample is meant to move the discussion beyond mere philosophical contrivances, to the actual explanatory practice of science: in the case under consideration, the explanation of face recognition.

1. Mechanistic versus functional models

In the explanatory practices of the life sciences, mechanistic models present a clear break from the old functional models that were once used, particularly in the cognitive sciences during the eighties and nineties, to explain a system's behavior. These older models explain the overall function of a system by decomposing that function into ever smaller sub-functions or -routines, and then showing how the overall behavior arises as a result of this functional organization, while remaining famously silent about the entities that were responsible ('realized') these (sub)routines.[4] Thus, for example, in psycholinguistics, speech production was explained by dividing the overall capacity into subroutines like conceptualization, formulation and articulation, while the subroutine of formulation was in turn decomposed into lexicalization and syntactic planning (Levelt, 1989). In this way, a rough explanatory model of speech production could be provided by means of a box diagram, which shows the different subroutines and their organization:

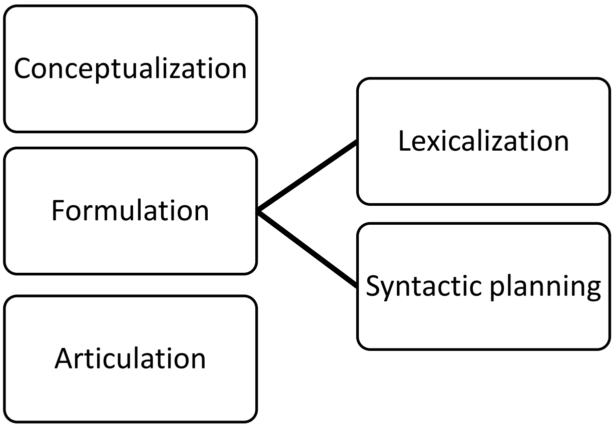

Figure 1: A functional model of speech production

Although popular in the 1970s and 80s, models like these are increasingly criticized for remaining silent about what (neural) entities realize a given (sub)routine. Of course, the sketch provided above is a very abstract, rough model of how the capacity to produce speech might be realized. Moreover, even if it is accurate in the activities it postulates, can it be really said to explain speech production if it has nothing to say about the neural entities responsible for all these activities? Only too often, researchers are at a loss about what is really behind the boxes in their diagrams. Now for some heuristic purposes, e.g. when we are just mapping out a certain capacity, this may be fine,[5] but if the original status of these boxes as mere placeholders is forgotten, they only serve to disguise gaps in our understanding ('boxology' is the derogatory term sometimes applied to such functional models).

In contrast, mechanistic explanations go beyond an abstract representation of the organization of activities, to cite the actual entities realizing these activities. A functional model might be useful for the purposes of prediction, mapping the input-output relation of a target system, but for a model to have explanatory powers, this is not enough. After all, using the Ptolemaic model of the heavens one can predict with some accuracy the location of certain celestial bodies in the night sky, but it does not explain why the planets move the way they do. In short, phenomenal adequacy is not sufficient for explanation, because *descriptive* accuracy is necessary.

This brings us to a further consideration: models vary in the degree to which they are accurate. Accordingly, Craver proposes a continuum on which a given model or mechanistic explanation can be placed, depending its degree of mechanistic plausibility (Craver 2006; the idea was first proposed in Machamer et al. 2000). The continuum ranges from speculative sketches, where a lot of the details are left out, to ideally complete descriptions, which identify every component in the mechanism. He distinguishes three developmental stages (Craver 2006 p. 361):

1) **How-possibly models**. These models are speculative conjectures about how a capacity might be realized: specifying a set of possible parts and activities that account for the behavior of the mechanism.
2) **How-plausibly models**. These are "…more or less consistent with the known constraints on the on the components, their activities, and their organization" (Craver 2006 p. 361).
3) **How-actually models**. These are ideal, complete descriptions of the model. They describe how the mechanism is composed and works in reality.

So far so good. However, does this continuum only represent the plausibility degree? Not according to Craver himself. He seems to relate it to the idea of explanatory power, for although he does state that "…how accurately a model must represent the details of the internal workings of a mechanism will depend upon the purpose for which the model is being deployed" he adds that "If one is trying to explain the phenomenon, however, it will not do merely to describe some mechanisms that would produce the phenomenon" (Craver 2006 p.361). It seems then that the further to the ideal description-side of the continuum a given model is placed, the more justified we are in calling a model an explanation. Why is this? The idea is that as one moves from how-possibly to how-actually models, the answers one gets allow a greater degree of control: "Deeper explanations show how the system would behave under a wider range of interventions than do phenomenal models" (Craver 2006 p. 358).

At first glance, one might be tempted to identify how-possibly models with functional explanations as I described them above. However, even the how-possibly models conjecture the existence of specified parts, although we might not even have any evidence that these parts exist. This would mean that purely functional models as described above are not explanatory at all, as they are deliberately silent about the realizer of a given function or sub-function (which, as you might recall, was what made functional explanations attractive to adherents of functionalism in the philosophy of mind). Like the Ptolemaic model of the heavens, these are merely "phenomenal models" which only succeed in mapping the input-output patterns of the mechanism in question.

However, there is an ambiguity. Craver seems to be talking about two things. On the one hand, he clearly talks about the accuracy of a model; i.e. the degree of correctness or truthfulness of a model. The phrases 'how-plausibly' and 'how-actually' make this plain. But there is also another sense in which the continuum is ordered: from abstract descriptions, to more detailed, complete descriptions. This is evident when he talks of a how-actually models as being a "…ideally *complete* description…" (Craver 2006 p. 360 my italics), or when he says: "Between sketches and complete descriptions lies a continuum of mechanism schemata that abstract away to a greater or lesser extent from the gory details…" (Craver 2006 p. 360). In effect, *Craver conflates plausibility with richness, and holds the presence of both to be necessary for a model to be of explanatory power.*

This cluster of ideas is already present when, in their 2000 article, Machamer et al. distinguish between what they call 'mechanism sketches' and 'mechanism schemata', where the former are abstract, incomplete versions of the latter, in that their "…entities and activities cannot (yet) be supplied…" (Machamer et al. 2000 p. 18). In other words, they are less rich in details about parts and operations, and only when such information has been added are they considered to be full-fledged mechanism schemata that are required for mechanistic explanation.

Nor are these authors alone in asserting that such a conflated notion of plausibility and richness is required for explanation. Glennan, for example, writes:

> The requirements for a model being a description of a mechanism place substantive constraints on the choice of state variables (such as the fact that state variables should refer to properties of parts), parameters and laws of succession and coexistence. The satisfaction of these (…) constraints is what accounts for the explanatory power of mechanical models (2005 p. 448).

while Bechtel claims that advancing a mechanistic explanation

> ...requires decomposing the mechanism into component parts and operations and localizing each operation in the appropriate part (Bechtel 2007 p. 176).[6]

To recapitulate then, it is reasonable to distinguish between mere phenomenal and genuinely explanatory models. In the literature on mechanistic models, there is a tendency to claim that what is required for a model to be of explanatory value, is that it specifies the actual mechanism responsible for the explanandum. However, we have seen that this is really a two-part job: the model has to provide a description of a the parts and operations of the mechanism, and it has to do so in an accurate, informed way. That is, the model should be rich in that it gives details about entities and activities, and these details should be plausible.

2. Plausibility and richness vary independently

However, at least conceptually, there are good reasons to keep these two features a model can exhibit separate. A higher degree of plausibility, for example, does not entail a higher degree of richness. It just means there is more evidence that the model is a true description. Likewise, models may be rich in detail and yet be plainly wrong. For one, though incorrect, the Ptolemaic model of the heavens is very rich in detail, postulating deferents and epicycles, fixed stars, nine celestial spheres etc. In a sense, the more details one gives, the greater the chances of error, and the more abstract a model is, the more likely it is to be right in the few assertions it does make (to the point of triviality). Let us hold on to Craver's terminology for plausibility, and introduce the terms how-abstractly, how-partially and how-concretely to cover richness. We now find ourselves with two continuums.

How-possibly → How-plausibly → How-actually

How-abstractly → How-partially → How-completely

Two features models can have ordered on two continuums

In principle, every combination is possible. There might be complete models that are only very loosely supported by evidence (how-completely/how-possibly), highly abstract models that nevertheless enjoy considerable support in the few assertions they do make (how-abstractly/how-actually), etc.

There is an additional complication however. Recall that mechanisms comprise both entities as well as activities. Both plausibility and richness, in all their respective strengths, can apply independently to entities and to their operations. This means that the conceptual space of possibilities we have so far explored is extended with yet another dimension. A model might contain a lot of information about a mechanism's parts while little about the operations they perform. Yet the little it says about the operations might be far more plausible then the detailed story it gives about the entities. Then again, in another model the degrees of plausibility and richness might more or less converge with respect to activities, while strongly diverging when it comes to the entities.

Anyway, it should be clear that at least conceptually, plausibility and richness can vary independently from one another, both with respect to entire mechanisms and within a particular mechanism. But of course, this is only part of the story: there remains the issue of explanatory power. By holding both plausibility and richness as necessary for explanation, one cannot make sense of the functional models described in the previous section. It is here that the cost of conflating these two notions is truly felt, as it hampers our understanding of the explanatory practices of the actual scientists themselves. In the next section, I will bring this point home by considering a functional model of face recognition.

3. Plausibility and richness in models of face recognition

Face recognition, or the capacity to spot faces from among other sensory data is a socially advantageous trait, as it enables us to make judgments about fitness, sex, health and emotional status of other individuals. Indeed, so highly developed and common is this capacity in humans, that it often argued to have a genetic basis (Wilmer et al. 2010; Zhu et al. 2010). In fact, we are so attuned to that faces sometimes the ability is triggered by certain features of non-face objects (e.g. distance and size ratios of rocky protrusions on a mountain surface). The idea that face recognition is a special case, i.e. requires its own explanation, separate from the general models for object recognition, is based primarily on evidence that the inversion of faces affects the ability to recognize faces more than it does other objects: the so-called *face inversion effect* or FIE (Yin, 1969).[7]

Early explanations of the capacity of face recognition relied heavily on functional analysis (e.g. Marr and Nishihara 1978, Rhodes 1985). Let us here consider the example of Bruce and Young's functional model of face recognition (Bruce and Young 1986).

They explained human face recognition in terms of a functional model (see figure 3), in which visual data of a presented face is structurally encoded to produce two different types of descriptions (view-centered and expression-independent), which in turn are analyzed in three different ways (analysis of facial speech and expression apply to the former, face recognition units analyze the latter). Choosing to remain silent about the details of the first two types of analysis, they ascribed to each face recognition unit a number of stored structural codes, each one describing a face that had already been seen before and was stored in memory.

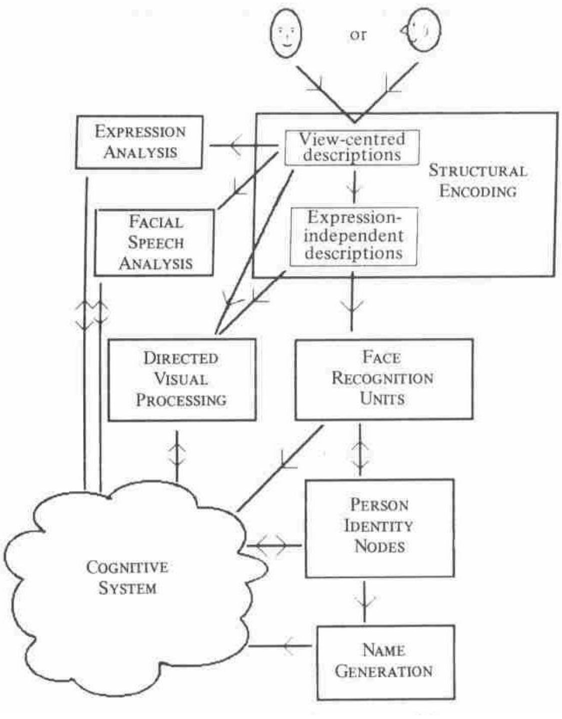

Figure 2: Functional model of face recognition (source: Bruce and Young 1986)

Here we have an example of a classic functional explanation. Bruce and Young said about their model: "…we are concerned almost exclusively with evidence in favor of functional components in the human face processing system, without regard to whether or not these are localized to specific areas of the brain" (Bruce and Young 1986 p. 306). In fact, they acknowledged the threat of boxology: "We recognize that the differences in the statuses of the arrows and the boxes used in models of this type are problematic" (Idem p. 311). However, speculative as this analysis may be, it is not simply fantasizing. There are reasons why the boundaries between the sub-functions have been drawn the way they are: "A 'box' represents any processing module, or store, which plays a distinct functional role, and whose operation can be eliminated, isolated or independently manipulated through experiment or as a consequence of brain damage" (Idem p. 311). For example, the reason to distinguish between subtasks having to do with person recognition (the 'person identity nodes' box) and with face recognition (the 'face recognition units' box) is that in experiments, it was found that face recognition can break down while person recognition remains intact (Hécaen 1981). Moreover, although generally not interested in the neural localization of the functions they proposed, Bruce and Young did used evidence obtained from experiments on people whose face recognition skills were impaired (prosopagnosia) to support their model (Idem p. 315). In fact, it seems that this functional model allows suggests ways to intervene upon the causal process for the benefit of experiment: just like Craver requires explanatory models to do, Bruce and Young's model shows "…how the system would behave under a wider range of interventions than do phenomenal models" (Craver 2006 p. 358).

So, where does this model fit on the continuums we considered in the previous section? With regard to plausibility, the researchers themselves admit it is speculative. However, as we have seen, part of the model was based on experimental evidence, so we can count this model as at least plausible to some degree, which in any case is all that is needed. The first condition is thus met. But what about richness? The model is rich in information regarding the mechanism's activities (analyzing, processing, encoding generating), yet it says next to nothing about what realizes these activities. In fact, *it is a how-abstractly model with regard to entities*. Here we can see the true cost of the confusion we mentioned in the first section: if we do not distinguish between plausibility and richness, and within the latter, between richness regarding entities and regarding activities, then it seems we are forced to say that this model does not explain human face recognition. However, it seems that by leaving out the details by which all the sub-routines are implemented, the model in fact highlights just those features of the mechanism that are explanatorily relevant.[8]

To conclude, this model is not like the Ptolemaic model of the heavens: it postulates operations and makes distinctions based on experimental evidence, and so is more than a mere input-output mimicking device. It provides information about what goes on between input and output, and does this in a plausible way. However, it does not provide information on the parts or entities of the mechanism, and in that sense, does not meet the condition of richness. If we were to follow Craver, we would have to discard Bruce and Young's model as 'merely phenomenal' and lacking explanatory power.

Conclusion

In this article, I have argued for two claims: first, that the literature on mechanistic explanations tends to confuse two features of models, plausibility and richness, and second, that this confusion can cloud our view on the practice of scientific explanation because *both* features are deemed necessary for a model to have explanatory power. Regarding the first claim, I have argued that plausibility and richness can vary independently from each other. Regarding the second, I have argued that traditional functional models constitute counterexamples, case in point being Bruce and Young's functional model of face recognition. These functional models go beyond merely input-output mapping to provide information on the operations of a mechanism that secures their status as explanations, yet they are neutral about the entities that perform the operations they stipulate.

The awkward conclusion that all these functional models have no explanatory power can be avoided if one distinguishes between plausibility and richness. While the former may be necessary for explanation, the latter is not, at least not with regards to a mechanism's entities.

References

- Bechtel, W. (2007). Reducing Psychology while Maintaining its Autonomy via Mechanistic Explanations. In M. Schouten and H. Looren de Jong (eds.), *The Matter of the Mind. Philosophical Essays on Psychology, Neuroscience, and Reduction*. Oxford: Blackwell Publishing.
- Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience.* New York: Routledge.
- Bechtel, W., & Abrahamsen, A. A. (2005). Explanation: A Mechanist Alternative. *Studies in the History and Philosophy of Biology and the Biomedical Sciences*, 36, 421-441.
- Bruce, V. and Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305-327.
- Craver, C. F. (2006). When mechanistic models explain. *Synthese,* 153, 355-376.
- Craver, C. F. (2007). *Explaining the Brain*. Oxford: Clarendon Press.

- Darden, L., & Craver, C. F. (2002). Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in the History and Philosophy of Biology and Biomedical Sciences* 33, 1-28.
- Fodor, J. A. (1981). Special sciences. In *Representations: Philosophical essays on the foundations of cognitive science* (pp. 127-145). Hassocks: Harvester.
- Gervais, R. R. (forthcoming). Some comments on the explanatory power of models in the cognitive sciences. *Proceedings of the LRR10 Conference on Logic, Reasoning and Rationality.*
- Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science,* 69 *(Supplement)*, S342-S353.
- Glennan, S. (2005). Modeling mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 443-464.
- Hécaen, H. (1981). The neuropsychology of face recognition. In G. Davies, H. Ellis and J. Shepherd (eds.) *Perceiving and Remembering Faces*. London: Academic Press.
- Kanwisher, N. G. & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions – Royal Society. Biological Sciences,* 361, 2109-2128.
- Levelt, W. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
- Machamer, P. K., Darden, L. & Craver, C.F. (2000). Thinking About Mechanisms. *Philosophy of Science,* 67, 1-25.
- Marr, D. and Nishihara, H. K. (1978). Visual information processing: Artificial intelligence and the sensorium of sight. *Technology Review,* 81, 1-23.
- Rhodes, G. (1985). Lateralized processes in face recognition. *British Journal of Psychology,* 76, 249-271.
- Tabery, J. G. (2004). Synthesizing activities and interactions in the concept of a mechanism. *Philosophy of Science*, 71, 1-15.
- Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K. and Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the Unites States of America,* 107 (11), 5238-5241.
- Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology,* 81, 141-145.
- Zhu, Q., Song, Y. Hu, S. Li, X., Tian, M., Zhen, Z., Dong, Q., Kanwisher, N., Liu, J. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology,* 20 (2), 137-142.

[1] As Machamer et al.'s classic definition has it, mechanisms are "…entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al. 2000). Note that in choosing the phrase 'are productive of', the authors do not (yet) specify whether the relation between the mechanism and the phenomena is one of causation, constitution or something else. Depending on what term you choose, some deep metaphysical problems concerning the causal status of higher level entities might arise. As these issues have no direct bearing on the points I wish to make in this paper, I will not take a stance on this issue.

[2] Accuracy should not be understood here as exact isomorphic correspondence between the model and the mechanism, but rather as a degree of similarity: the term plausibility is partly chosen to reflect this.

[3] The idea for such a continuum was already present in Machamer et al. 2000.

[4] A fact which, of course, made them popular with philosophers of mind arguing for the so-called autonomy of the special sciences (see for instance Fodor 1981).

[5] A remark of this sort is made in Machamer et al., where it is said that mechanistic explanations typically start by providing a 'mechanism sketch', which is an "…abstraction for which bottom out entities and activities cannot (yet) be supplied or which contains gaps in its stages. The productive continuity from one stage to the next has missing pieces, black boxes, which we do not yet know how to fill in" (Machamer et al. 2000 p. 18).

[6] The word 'component' in the jargon of the mechanists both refers to the entities of a mechanism, and its activities (Darden 2006).

[7] This so-called face-specificity hypothesis is not without its critics however. For a survey of the evidence and the literature regarding this issue, see Kanwisher & Yovel (2006).

[8] In fact, in technological contexts, we often want our models to be rich in details about activities, rather than entities, let alone general plausibility. When it comes to artificial systems, we demand of our models that they duplicate the function of a natural system, not that they accurately describe the way in which they are implemented: in such cases, performance trumps accuracy (Gervais, forthcoming).