MODELS AND EXPLANATION

Elanor Taylor

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Philosophy.

Chapel Hill
2008

Approved by:

William Lycan
Dorit Bar-On
Marc Lange

ABSTRACT

Elanor Taylor
Models and Explanation

(Under the direction of William Lycan, Dorit Bar-On and Marc Lange.)

How can a description of a thing act as an explanation?

This is a question that hangs over much of the philosophical enquiry into the role of models in science. In this paper, I attempt to address this question by examining a particular use of models in cognitive science, dynamic systems theory, and in particular the Haken-Kelso-Bunz mathematical model of phase transitions in finger movements. I defend a view on which this model provides resources for what I call "systems explanation", where a systems explanation is a scientific explanation that proceeds by giving a profile of an interactive system that gives rise to the target phenomenon.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

How can a description of a thing ever act as an explanation of it?

This is a question that hangs over much of the philosophical enquiry into the role of models in science. In this paper, I attempt to address this question by examining a particular use of models in cognitive science, dynamic systems theory, and in particular the Haken-Kelso-Bunz model of finger phase transitions. I defend a view on which this model, which is a set of differential equations, provides resources for what I call "systems explanation", where a systems explanation is a scientific explanation that proceeds by giving a profile of an interactive system that gives rise to the phenomenon that you are trying to explain.

Chapter 1: Dynamic systems theory in cognitive science

Dynamic Systems Theory is a branch of mathematics that can be used to model the evolution of the behaviour of systems through time. As an approach to cognitive science, DST methods stand apart from more traditional computational approaches that understand cognitive phenomena in terms of "rules and representations", positing internal representations and rules that operate over them[1]. There's a significant amount of debate about whether or not representational and non-representational approaches are compatible, which has resulted into the development of a number of hybrids that combine DST with representational approaches[2]. I will discuss this in passing debate later on, but suffice to say for the moment that DST approaches differ from cognitive science traditionally-construed in that they deal with phenomena in terms of the activity of whole systems, often in interaction with environmental factors, rather than in terms of the manipulation of internal representations.

I will go on to outline a classic example of dynamic systems theory in action – the HKB model of phase changes in finger movements – but before doing so I will draw some preliminary distinctions.

I will use the phrase "dynamic systems theory" for the mathematical methodology described in rough terms above. Dynamic systems theory is, on this use, an approach to modelling and explaining phenomena that views such phenomena as whole systems in time, using a given set of concepts (to be outlined in later discussion) and a set of mathematical tools. An interesting feature of dynamical approaches is that they offer a geometric picture of the development of the target system over time, by drawing out what is known as the *phase space* of

[1] See the section on explanation in cognitive science for a fuller explication of these differences and of the "rules and representations" paradigm.

[2] See for example Clark 1997a and Eliasmith 2003.

the system and enabling the tracing of paths and patterns in this phase space. Dynamical models themselves are sets of equations.

Some of those who advocate the use of DST in cognitive science also hold a strong position on the practice of cognitive science and on the nature of cognitive phenomena, called *dynamicism*. Dynamicism in its strongest form is the view, expressed by Port and Van Gelder as the *dynamical hypothesis*, that *"Natural cognitive systems are dynamical systems and are best understood from the perspective of dynamics*"[3]. Obviously, this needs some unpacking; Port and Van Gelder define dynamical systems roughly as *"systems with numerical states that evolve over time"*[4]. They go on to expand on this view, claiming that dynamical systems are "state-determined systems", where state-determined means that the system is such that "*its current state always determines a unique future behaviour"*[5]. This claim could be interpreted in a number of ways, some of which would make the notion of a state-determined system trivial, but the most plausible version is that a state-determined system is an interactive set of aspects of the world such that the future states of the system are always uniquely determined by the current state in accordance with some rule. The rule may change as a function of time if the system is affected by external factors (so the set can either be self-contained or not) and such factors are known as *parameters* in DST parlance. Some critics of dynamicist approaches to cognitive science argue that adopting the method shows very little about the phenomena being modelled; if anything can be modelled dynamically then there is nothing very interesting about it. Port and Van Gelder answer that *not* everything can be modelled as a dynamical system, and showing that cognitive systems *are* dynamical systems requires a substantial amount of empirical work. This is a controversial claim, and so there is a substantial difference between the philosophical commitments of strong dynamicism and the theoretical commitments of dynamic systems theory.

[3] Port and Van Gelder ed. 1995 pg 5

[4] Ibid pg 5

[5] Ibid pg 6

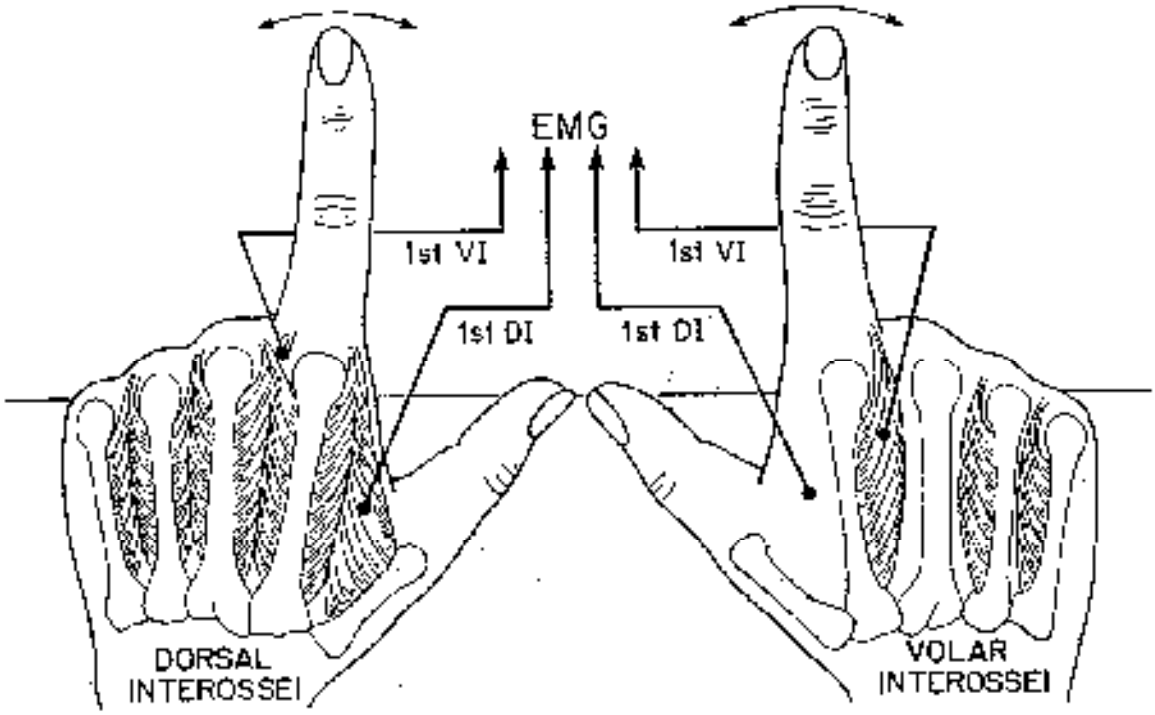
Overall, then, we have the following definitions on the table:

- Dynamical system: Any state-determined system with a numerical phase space and a rule of evolution specifying trajectories in this space.
- Dynamic Systems Theory (DST): The branch of mathematics that studies state-determined systems.
- Dynamic Systems Theory in Cognitive Science (DST-C): An approach to cognitive science that uses the tools and resources of DST to study natural cognitive phenomena.
- Strong Dynamicism: The belief that natural cognitive systems are dynamical systems and are best understood from the perspective of dynamics.

DST-C in action: The Haken-Kelso-Bunz Model of Phase Changes in Human Finger Movements

This is one of the classic examples of DST-C at work. I bring it up here not only for its historical importance but also for its simplicity. Some critics of DST-C have focused on the fact of this simplicity, arguing that so far only "low-level", fairly simple cognitive phenomena have been fully modelled in the DST framework, and that this reflects the limitations of the approach. This question about the purview of DST-C is important because if it turns out that it is really only useful as a way of modelling, say, motor control, but its tools are of no use when tackling more complex cognitive phenomena, then the revolutionary aims of strong dynamicism start to look a little shaky. For the moment, however, the HKB model of phase transitions in finger movements will provide a useful case study because of the simplicity of the model and of the target phenomena.

The finger-phase experiment was inspired by the phenomenon of gait transition. The simple phenomenon at its heart is the fact that, when a person moves their two index fingers from left to right across the same plane and at the same frequency, there are two possible "gaits" or phases of movement. In-phase movement is where the equivalent muscles of each hand contract in time, and anti-phase movement is where one contracts and the equivalent opposing muscle expands at the same time. After a certain frequency of oscillation is reached, however, the movements converge onto in-phase no matter which phase they started out in. Kelso reports pondering on the possibility of an analogy between gait transitions and non-equilibrium phase transitions, which, as he puts it, are *"the simplest form of self-organisation known to physics"*[6].



Having established this phase-shift in human hand movements in a laboratory setting, Kelso and colleagues sought to model it as a phase transition. As mentioned above, it turned out

[6] Ibid pg 46

that the phases of movement showed a steady pattern until a critical value was reached in the frequency of oscillation. The model of the phenomenon, known now as the Haken-Kelso-Bunz model after Kelso and his collaborators, is a set of equations that describes the system's evolution through the relative *phase space,* where the phase space is a representation of all the possible states of the system and how these evolve in time (Kelso described this as being where a dynamical system "lives").

I'll take a brief foray into some DST language, as it will be useful for understanding the HKB model. Some dynamical systems have phase spaces with *attractors*; these are states of the system that are simply more popular than others, towards which the system converges. A *bifurcation* occurs when a control parameter reaches a critical point that changes the attractor (note that this is equivalent to the physicist's nonequilibrium phase transition). In the finger movements example, the system at standard frequency has two attractors – the in-phase and anti-phase movements. At the critical point at which the control parameter changed, i.e. the frequency of oscillation reached a key value, a bifurcation occurred whereby a system with two attractors became a system with one. With these definitions on the table, we can understand what Kelso means when he says:

> *For now, the task is to postulate the simplest mathematical function that could accommodate space-time symmetry, bistability and the observed bifurcation diagram in the walking fingers experiments*[7].

Kelso et al then mathematically modelled the gait transition behaviour. This involved developing a set of equations that described the state space of the system, the possible developments in time of relative phase in relation to the control parameter. Mathematics aside, for the purposes of the following discussion we need take away only one central point - it is the equations themselves that are the model.

To sum up, DST-C takes a set of behaviours and examines this behaviour as the activity of a whole system. Patterns in the behaviour are modelled using mathematical tools that provide a

[7] Ibid 1995 pg 54

description of the phase space of the dynamical system, and allow us to pick out key features of those patterns (such as bifurcations and attractors). The HKB model just is this mathematical description

Chapter 2: Scientific Explanation and Cognitive Scientific Explanation

In order to ask whether DST-C models are explanatory we need some idea of what a scientific explanation is. Ideas about what counts as an explanation vary across philosophy of science and philosophy of cognitive science. I don't intend to defend a particular position on the nature of scientific explanation here, though I do adhere to a general view of scientific explanation as a family resemblance concept[8]. Instead, I'll discuss a series of features that we could expect from such an explanation, and go on to briefly discuss styles of explanation in cognitive science, running through the traditional views on explanation in cognitive science with the aim of providing some background against which to investigate the explanatory role of DST-C models.

Features of scientific explanation

An obvious fact to note is a difference in medium between any kind of explanation and the DST-C models. Explanations are typically linguistic entities. We *give* explanations, in language. If this was part of the criteria for scientific explanation, the DST-C models would fall at the first hurdle. We can distinguish between the answer to such a question, however, and the means of that answer's communication. Explanations may be answers to why-questions, but there is no need to equate the answer itself with its linguistic expression.

[8] Thanks to Bill Lycan for this suggestion.

One general feature of scientific explanation has already been mentioned – they are often *answers to "how" or "why" questions*[9]. Clearly this requires significant pragmatic caveats; a why question in the language of, say, biochemistry, is a request for a biochemical explanation, whereas a why question in the language of history is a request for a historical explanation. Salmon points out that many requests for scientific explanation in fact take the form of a "how-possibly?" question, for example in a case where we ask how it was possible for a cat to land on its feet after falling from a certain angle[10]. Van Fraassen argues that all requests for scientific explanation can be re-phrased as why-questions, claiming that context will be all that makes the difference between explanatory and descriptive language. I will leave the question of whether all explanations are answers to why-questions aside for now – suffice to say that being an answer to a how-or-why-question is an item on the list of prototypical features of scientific explanation.

Another item on this list is *uncovering mechanisms*. This can be understood in many different ways, but the following quote from Bechtel and Abrahamsen captures the general idea:

> *A mechanism is a structure performing a function in virtue of its component parts, component operations and their organisation*[11].

A mechanistic explanation thus proceeds by giving an account of the structure that gives rise to some phenomenon, discussing the roles of the various parts that make it up. I will discuss the notion of mechanistic explanation further in later sections, in particular expanding on the requirement that a mechanistic explanation gives an account of some phenomenon in virtue of a decomposition of the parts of the structure that give rise to it.

The final general feature of scientific explanations that I will discuss is that they often attempt to *isolate a single causal factor* responsible for a given phenomenon. I am wary of saying that such explanations isolate "the cause" of a phenomenon, or even "a cause". Suffice to

[9] A view that has been most famously explored by Van Fraassen. See Van Fraassen 1980

[10] Salmon 1989 pg 137

[11] Bechtel and Abrahamsen 2006, pg 3

say that there is a sense in which we expect scientific explanations to enable us to understand how or why something came about, and this style of explanation does so by picking out the individual factor thought to be responsible for the phenomenon at hand. I'll discuss this idea of single-factor explanation in more detail in my second section, using Salmon's Statistical Relevance model of scientific explanation as an example, and will explore the commitments involved in taking such explanations to be causal.

I will therefore take the following three features to be prototypical of scientific explanations:

1) Answer a how-or-why question.

2) Uncover a mechanism.

3) Isolate a, or the, relevant causal factor.

There are a number of questions about scientific explanation that apply across such models. Salmon makes the distinction between ontic and epistemic notions of explanation, where ontic approaches aim to pick out "things in the world" that make something the case, and epistemic approaches aim to, say, make phenomena more probable. I would argue (against Salmon, who takes a different view) that such differences can apply across the three notions of explanation presented here. A single-factor approach could present some factor as causally, or "ontically" responsible for the phenomenon at hand, but could also present the same factor as a feature that makes the target phenomenon more likely. These questions apply across the features I have presented here. For most of this discussion I will assume an ontic model, but the views presented could easily be reinterpreted on an epistemic model.

Explanation in cognitive science

Traditionally, explanation in cognitive science falls into three main categories – symbolicist, connectionist and dynamicist. These categories can be misleading – they are in no

way exhaustive or exclusive, and, as I will discuss later, there are many different ways of carving out the explanatory categories of cognitive science. It is useful to outline these three categories just now, however, as a route into debates about explanation in this area[12].

Symbolicism is the background architectural assumption of traditional artificial intelligence and cognitive science. It trades off the notion that cognitive processes involve internal representations being operated over by rules – otherwise known as the "rules and representations" paradigm. A symbolicist explanation of some cognitive phenomenon would therefore typically offer an account of that phenomenon in terms of mental representations and the rules that operate over them. Symbolicist explanation operates against the backdrop of the computational metaphor of mind, whereby the mind is metaphorically understood, for the purposes of cognitive scientific investigation, as displaying something like a von Neumann architecture.

The symbolicist paradigm was followed by connectionism, which also operates on an analogy, this time of a weighted network of nodes. Connectionism provided an empirically and perhaps more biologically plausible view of mental architecture (accounting for phenomena like graceful degradation[13]). Connectionist networks are not necessarily anti-representationalist, however – the inner symbols of classical artificial intelligence can be thought of as quite happily being *realised* on a connectionist network. A connectionist explanation would typically offer some account of the target phenomenon in terms of patterns of activity in a weighted network, and particularly the network "learning" through experience.

Dynamicism is the position that we are in the process of exploring – dynamicist approaches seek to model cognitive processes by tracing the evolution of the behaviour of systems through time using a variety of mathematical tools. Note that, just as connectionist networks can realise

[12]This section owes a lot to the introductory material in Eliasmith 2003.

[13] Graceful degradation is the phenomenon whereby noisy input or destruction of some unit in a system leads to a gradual dropping-off in efficacy of function, rather than complete collapse. This is thought to be a feature of, e.g. the erosion of memory in some mammals. Connectionist networks are more successful than classical networks at handling such phenomena.

representations, so can a dynamicist system, even if the dynamicist approach to explanation doesn't posit a symbolic architecture. This is an issue that can be effectively addressed by thinking about representations at different levels of realisation. Note, however, that many proponents of dynamicism reject the computational view of the mind. For instance, the "strong dynamicists" I referred to earlier claim that the best way to understand cognitive phenomena is from the perspective of dynamics, which commits them to a rejection of the computational model of cognitive scientific explanation.

Symbolicism, connectionism and dynamicism, as described, make up the three main styles of cognitive scientific explanation. Dynamicism differs from the first two in that it relies less on abstract computational metaphors and doesn't posit internal representations, although representationalism isn't incompatible with the adoption of dynamicist approaches to explanation. This is the traditional picture of explanation in cognitive science. More recently, however, a strand of explanatory pluralism has developed in philosophy of science, whose proponents include Clark and Eliasmith.

Clark argues that we need three types of explanation to do effective cognitive science – componential, catch and toss and emergent. Componential explanation is explanation in terms of a thing's component parts, catch-and-toss explanation works in terms of interactions between a thing and its environment, and emergent explanation takes the phenomena of interest as highly interactive whole systems, drawing little or no distinction between an object and its surroundings. Clark argues that DST is a promising tool for the purposes of emergent explanation, but that otherwise these three categories cut across the traditional triad of approaches to cognitive science outlined above. As he puts it:

> *To explain such heterogeneous phenomena, the theorist should be willing to exploit multiple kinds of explanatory tools, ranging from analyses that criss-cross the organism and the environment, to ones that quantify over multiple inner components and complex connectivities, to ones that isolate components and offer a functional and representational commentary on their basic roles*[14].

[14] Clark 1997a pg 127

Surveying the literature on styles of explanation in cognitive science, it is clear that a place has already been made for the notion of DST models as explanations. Eliasmith included dynamicism in his classical triad of cognitive scientific explanation, while Clark puts DST-C forward as the best candidate for the emergent explanation that he argues is required to address complex phenomena. Having surveyed some different approaches to scientific explanation and to cognitive scientific explanation, it is clear that these approaches involve different sorts of tasks. There is an important extra element in cognitive scientific explanation, which is the particularly cognitive character of the phenomena at hand. The aim of a cognitive scientific explanation is to take some phenomenon and explain it as the product of a thinking, or cognising, thing, to effectively ask "*what would a cognising thing have to be like to produce such a phenomenon?"* The aim of a scientific explanation is broader, asking instead "*what would have to be true of the world for such a phenomenon to obtain?"* Scientific explanations are also often thought to have interesting connections with theories and laws of nature, whereas this status doesn't extend to cognitive scientific explanation. An interesting feature of DST-C is that the explanatory approach it offers answer the second question more successfully than the first. There need be nothing specifically mental about the phenomena that a DST-C explanation targets, which partially accounts for its controversial status as an approach to cognitive science.

I will try to look more closely at the *kind* of explaining, if any, that DST-C models do, against the backdrop of this survey of ideas about scientific explanation and cognitive scientific explanation. Many philosophers of cognitive science seem happy to accept the idea that DST-C models explain in some sense or another. The question I am interested in is in virtue of what do they perform this task. The rest of this discussion, then, will be focused on this question – explaining how DST-C models explain.

Chapter 3: Isolating single factors and uncovering mechanisms

In this section, I'll address the relationship between DST-C models and the key features of scientific explanation discussed above. I'll argue that DST-C models fail to answer scientific why-questions in the two ways discussed, isolating a single factor or uncovering a mechanism.

For simplicity's sake, I'll focus on the DST-C model already discussed, the HKB model of phase transitions in finger movements. This model provides a law-like summary of patterns in the behaviour of the two fingers, observed as a whole system developing over time. Could this model be an adequate answer to a why-question?

There are a number of likely candidates for the relevant why-question in this case. I'll take the following as a central example:

Why do the finger movements change phase in this way?

The issue at hand is whether or not the model acts as an answer to this question, or indeed to any other. Of course, we need not expect a scientific explanation to come in linguistic form. It may be that DST-C models provide resources for scientific explanations – facts, patterns etc – without acting as explanations themselves, but their not acting as explanations in themselves won't be a matter of their being linguistic or not. Above, I outlined three general features of scientific explanations – being an answer to a why-question, isolating relevant causal factors and uncovering mechanisms. I will presuppose the format of the first dictating the second two, so in the following discussion I will look into how the model could answer a why question either by isolating a or the relevant single factor or uncovering a mechanism.

First of all, I would like to make clearer the distinction between isolating a or the relevant single factor and uncovering a mechanism. As an example of the first, I'll look briefly at

Salmon's Statistical-Relevance (henceforth S-R) model of scientific explanation. I don't intend to defend it as *the* definitive account of scientific explanation, or indeed of causal explanation, but instead to show the *kind of thing* an explanation that isolates a or the relevant single factor is.

Salmon describes working towards the S-R model of scientific explanation as an attempt to provide an alternative to Hempel's Inductive-Statistical model that *"would be based on statistical relevance rather than high probability"*[15]. The basic idea of a statistical relevance explanation is that statistically relevant information is explanatory, where statistical relevance is understood as follows:

> *Given some class or population A, an attribute C will be statistically relevant to another attribute B if and only if P(B|A.C) ≠ P(B|A) — that is, if and only if the probability of B conditional on A and C is different from the probability of B conditional on A alone.*[16]

Salmon took the canonical form of the request for a scientific explanation as: *Why does this (member of the class) A have attribute B?"*[17] So, for example: *Why did Jones, who is a member of the class of people who have strep infections, recover quickly?*

To put it in very simple terms, the S-R model of scientific explanation proceeds by isolating a class of things of interest, and dividing this class into relevant partitions, where "relevant" indicates that the probability of the members of the class having some attribute is different across the partition. Remember that the request for an S-R explanation involves asking why a member of a given class has a given attribute. The S-R explanation will point out that this member is a member of a certain partition of the class, a partition in which there is a higher probability that the members have the relevant attribute than in the rest of the class. Given the assumption that statistical information is explanatory, this fact about partition-membership is an explanation of the target phenomenon.

---

[15] Salmon 1989 pg 62

[16] Woodward 2003

[17] ibid

To return to the strep case, the following quote from Woodward gives the S-R explanation for Jones' quick recovery:

> *… suppose we want to construct an SR explanation of why x who has a strep infection = S , recovers quickly = Q. Let T (-T) according to whether x is (is not) treated with penicillin, and R(-R) = according to whether the subject has a penicillin-resistant strain. Assume for the sake of argument that no other factors are relevant to quick recovery. There are four possible combinations of these properties: T.R, -T.R, T.-R, -T.-R, but let us assume that P(Q|S.T.R) = P(Q|S.-T.R) = P(Q|S.-T.-R) ≠ P(Q|S. T.-R). That is, the probability of quick recovery, given that one has strep, is the same for those who have the resistant strain regardless of whether or not they are treated and also the same for those who have not been treated. By contrast, the probability of recovery is different (presumably greater) among those with strep who have been treated and do not have the resistant strain.*
>
> *In this case [S. (T.R v -- T. R v-R.-T)], [S.T.-R] is a homogenous partition of S with respect to Q. The SR explanation of x's recovery will consist of a statement of the probability of quick recovery among all those with strep (this is (i) above), a statement of the probability of recovery in each of the two cells of the above partition ((ii) above), and the cell to which x belongs, which is S.T.R ((iii) above).*[18]

There are further details of the account that are unnecessary to go into for the sake of this discussion (although it has many interesting features, including the screening-off relation and the handling of low-probability events). I take the S-R model of scientific explanation as an example of the kind of explanation that attempts to uncover a or the single factor. There are interesting questions about whether or not S-R explanations uncover *causal* factors, but for the case at hand it is only the structure of these explanations that I am interested in. The S-R model of explanation involves a search for something, an isolatable factor, that is relevant to, or in some sense responsible for, the phenomenon at hand. Whether this thing is cashed out as explicitly causal or not is beside the point; a general sense of its being "responsible for" the given phenomenon is enough. As an aside, note that taking such factors to be causal would require a very broad notion of cause, including background conditions and the absence of given attributes, and that this consideration tells against a purely causal reading of the S-R model[19]. Later on I will argue that

[18] ibid

[19] Thanks to Jesse Prinz for pointing this out

the DST-C model neither explains in this way nor provides resources for scientific explanations of this general form.

Uncovering a mechanism is a different kind of task. There are a number of views in the literature on mechanisms in explanation about how this uncovering should proceed, but it is typically understood in terms of decomposing the structure responsible for the phenomenon into well-understood constituent parts. Bechtel and Abrahamsen, for example, give the following account of a mechanism:

> *A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organisation. The orchestrated functioning of the mechanism is responsible for one or more phenomena*.[20]

And in a different text, Bechtel et al say the following of mechanistic explanation:

> *… what is crucial to such explanations is the decomposition of the actual system into component functions and component parts. It is by identifying the parts of the system, what they do, and how they are organised to work together, that one explains how a mechanical system performs its operations*[21].

Bechtel and Abrahamsen give their definition of mechanism in the context of a discussion of explanation in cognitive science and neuroscience, but they defend it as a thesis about scientific explanation in general. For the purposes of this discussion I'll assume this definition, whereby giving a mechanistic explanation involves explaining some phenomenon in virtue of the activity of the parts of the structure that gave rise to it[22].

---

[20] Bechtel and Abrahamsen 2006, pg 3

[21] Bechtel et al eds 2001, pg 12

[22] There are other notions of mechanistic explanation, some of which aren't committed to this idea of decomposition. For instance, the mechanical models offered by Maxwell were thought to support mechanistic explanations, but consisted of basically giving some story about how a phenomenon might have come about. I have chosen to concentrate on the decompositional notion of mechanistic explanation put forward by Bechtel and others first because it is more representative of contemporary views about mechanistic explanation in science, and second because it is interestingly different from other styles of scientific explanation in a way that it isn't clear the 19th century models are.

For a very simple example of a mechanistic explanation, think of the closing of the human jaw – this can be understood in terms of the functioning of its components and their interactions[23]. Breaking the movement of the jaw down into the components responsible for it and the roles that they play is just to give an account of the behaviour of the structure in virtue of its component parts. A more complicated example, the mechanistic explanation of carbohydrate metabolism, is given by Bechtel and Abrahamsen:

> *This is explained by decomposing the responsible mechanism into various enzymes (parts) that catalyze intracellular biochemical reactions (operations) in molecular substrates (another kind of parts)… The complete set of reactions is known as the Krebs cycle…The account can be completed by describing the spatiotemporal orchestration of the organised components in real time, that is, their dynamics*[24].

Note that there is a dynamical element to this particular mechanistic explanation, but that the explanation proceeds by decomposing the parts of the system, with the dynamical element coming in as part of the explanation of the nature of the interactions between the parts.

How, then, do these two ways of answering a scientific why-question apply to the HKB model of phase transitions in finger movements? I will argue that the HKB model neither provides a mechanistic explanation of the phase transitions in finger movements, nor provides resources for one, and that it also doesn't explain by isolating a single factor.

The HKB model captures patterns in the activity of a system. It doesn't break the activity of the system down into the activity of its components, whether you take "components" narrowly to mean just object-like (and thus traditionally "mechanical") entities, or broadly to include relational variables like frequency of oscillation or relative phase. The model doesn't break the behaviour down into components, nor does it break the object-like structure responsible for the behaviour down into components, and it doesn't provide resources for such an explanation. It therefore neither offers a mechanistic explanation nor offers resources for one.

---

[23] Thanks to Marc Lange for the example

[24] Bechtel and Abrahamsen 2006 pg 4

You could argue that the HKB model *does* offer a mechanistic explanation, in that the equations capture the behaviour of the components of the system, and so the equations' components facilitate the decomposition required for a mechanical explanation[25]. For example, if we had a set of equations describing some physical activity where the activity could be understood in terms of the contributions of, say, friction in one area and gravity in another, and the influence of these forces could themselves be represented in further equations. Where the original equations could be broken down into further equations that capture the role played by different physical forces with respect to the phenomenon of interest, a mechanistic explanation is made available in virtue of the decomposition of the equations. I would argue in response that the variables involved in the HKB model aren't decomposable in this way. Looking back to the HKB example, the variables represent features like frequency of oscillation of the fingers and the relative phase of the finger movements (their movements in relation to each other). These variables aren't decomposable into further physical forces "responsible" for the target phenomena, and so the HKB equations don't offer this kind of decomposition.

Rather than offering an account of the system in terms of the activity of its components, the equations describe the relations between those components. The model gives us an account of how the components of the system work together, but it doesn't give us an account of what any of those components "do". This therefore isn't enough for the kind of componential breakdown involved in mechanistic explanation, even on a broad understanding of "component" that includes variables like frequency of oscillation and relative phase, as componential explanation requires an account of the activity of the parts of a structure individually as well as their relations to each other. Note that mechanistic explanations can be given in purely relational terms – we don't need to assume that relational *means* non-decomposable to see that the HKB model isn't decomposable. For instance, consider a case where some highly complex purely relational phenomenon is broken down into simpler relational phenomena. This could perhaps, on Bechtel's

[25] Thanks to Marc Lange for the objection.

account, count as a mechanistic explanation (although the example would have to be substantially fleshed out). The fact that the equations capture relations between variables isn't what makes the HKB model non-mechanistic. Rather, it is the fact that the HKB model *only* describes relations which means that it provides no scope for the decomposition required in mechanistic explanation.

The HKB model provides a description of the phenomenon in question in the language (and conceptual framework) of Dynamic Systems Theory. Although the model itself doesn't provide an explanation, the model and the theory together provide *resources* for what I'll call "systems explanation"[26]. The systems explanation response to the question *"Why do the finger movements change phase in this way?"* will be as follows: *The finger movements change phase in this way because the fingers form a dynamical system, the activity of which is captured by this set of equations.*

What we have in the theory-model pair that we don't have with just the model is a way of interpreting the model. Understanding the model in light of the background theory is like taking a certain kind of explanatory stance towards the phenomenon – you are understanding it as a dynamical system, and having done so, the category of systems explanation becomes salient to it. I'll develop this notion of a systems explanation further in my final section.

Neither the model, nor the model-theory pair, acts as the kind of scientific explanation that isolates a or the relevant single factor. I presented Salmon's S-R model of scientific explanation as an example of the kind of scientific explanation that proceeds by attempting to find some single factor responsible for the phenomenon of interest. The systemic nature of the phenomenon in question in the HKB case rules this out; because of its systemic nature, there won't be a single relevant factor to find. There is a single fact about the phenomenon of interest that might look like the relevant causal factor – the point at which the phase transition occurs – but note that in the

[26] Cummins presents an idea of systems explanation, but this is an entirely different view. See Cummins 1975.

question above this is the very thing that we are trying to explain. The HKB model doesn't fit into even the broadest model of single-factor explanation.

Overall, then, the HKB model (and DST-C models more generally) provide a description of some phenomenon in the language of dynamic systems theory. Together, the model and the theory provide resources for systems explanations, but not for explanations that either isolate a or the relevant single factor, nor for explanations that uncover mechanisms. I will now go on to expand on this positive view, outlining the place of systems explanation in the domain of scientific explanation.

Chapter 4: Systems Explanation, Mediators and Counterfactuals

In this final section, I'll lay out my positive view of the explanatory role of DST-C models as providing resources for systems explanations, before going on to look at two alternative accounts of the explanatory role of DST-C models, Morgan and Morrison's idea of models as mediators and Clark's claim that DST models are explanations in virtue of their ability to "illuminate counterfactuals".

The positive account

I'll put forward the positive account of the explanatory role of DST-C models, before going on to argue for it and to flesh out the account further. The role of the models is as follows:

- Dynamic Systems Theory provides a conceptual framework to understand the target phenomenon within.
- DST-C models provide descriptions of that phenomenon within the framework of Dynamic Systems Theory.
- Together, the theory and the model provide resources for systems explanation of the given phenomenon, where a systems explanation is an answer to a why-question that gives an account of the systemic behaviour responsible for the target phenomenon.

Take the HKB case study. The target phenomenon is the predictable nature of the finger movements. Dynamic Systems Theory provides us with a conceptual framework within which we can understand the phenomenon as the activity of a system. The model describes the phenomenon in terms of the framework, giving us a picture of the activity as forming patterns of certain kinds,

developing through a given state space over time. Taking the target phenomenon, the explanandum, if you like, as before: *Why do the finger movements change in this way?* The answer will be as follows, given the DST-C model and the background theory: *The finger movements form a system with the state space described in the model. The point at which the finger movements change is a bifurcation point – the system changes from having two attractors to having one.* You might argue that this isn't an explanation at all, that there is a further question to ask about *why* this point is that point at which the system changes. This is where the notion of a systems explanation comes in; understanding some phenomenon as systemic means seeing the request for a mechanistic or single-cause explanation as inappropriate.

Understanding a phenomenon as systemic involves the assumption that interactions between the relevant variables will vary reciprocally. A system will, on this account, display interactive, reciprocal dependence between variables. For example in the HKB case, the speed of one finger's movement will dictate the speed of the other's; they co-vary reciprocally (remembering from the details of the experiment that subjects were asked to move their fingers in time with a metronome). Similarly, the bifurcation point and the frequency of oscillation will co-vary reciprocally. This is not a claim about the *direction* of causation – the bifurcation point arises out of the activity of the system, but the reciprocal dependence is intended to capture the idea that there will be no change in the one without a change in the other. Alternatively, given that changing a flagpole's length will change the length of its shadow, but changing the length of the shadow will not change the length of the flagpole, the flagpole and the shadow do *not* form a system. If the lengths did co-vary reciprocally then it would be appropriate to use a form of systems explanation to track the relations between those variables, but as it stands, without the assumption of reciprocity between variables a systems explanation isn't appropriate. This isn't intended to be an exhaustive characterisation of a system; indeed, it is an open question whether systems form a kind, or whether there will be a fact of the matter about whether some phenomenon is systemic. I will restrict my claim, then, to the following: understanding some

phenomenon as systemic brings with it the assumption that the relevant variables will co-vary in a reciprocal way. The justification for labelling that phenomenon as systemic will then be a matter of justifying this expectation of reciprocity.

Why, then, should we accept "systems explanations" explain anything? Because they perform an analogous role, relative to the phenomena they explain, that traditional (single-factor or mechanistic) scientific explanations do. In previous discussion, I pointed out that explanations that aim to isolate a or the relevant single factor proceed by picking out some feature responsible for the target phenomenon, and that recognising this explanatory structure doesn't impose a particular position on the role of cause in that explanation. An explanation based on the HKB model and the resources of dynamic systems theory will perform an analogous task. In giving a systems explanation, you are effectively giving a profile of the system that gave rise to the phenomenon, which is analogous, though not identical, to the isolation of relevant factors that I take as a central, prototypical feature of scientific explanation[27]. The relevant similarity is the idea of uncovering, or giving an account of what is responsible for the target phenomenon, whichever notion of causation you use to cash out this idea of responsibility. Due to the expectation of reciprocal interactions between variables involved in understanding some phenomenon as a system, providing a model that tracks relations between variables, rather than decomposes the phenomenon into parts or isolates a single causal factor, performs this task.

You could argue that, without either uncovering a mechanism or isolating a single factor, you have not successfully given a scientific explanation, that there is nothing explanatory about pointing to a description of a system. I would say in response, once you have decided that some phenomenon is systemic, you have effectively decided that an explanation that tracks relations between the relevant variables will give the best account of the phenomenon you are interested in. Deciding that something is systemic is to decide that there is a reciprocal dependence relation

[27] As mentioned previously, I adhere to a broad, "family resemblance" notion of scientific explanation, whereby uncovering mechanisms and relevant factors are prototypical features of scientific explanation but aren't necessary to it. I owe this idea to Bill Lycan.

between the relevant variables, and that interdependent relation can only be brought out at the systems level. The systems explanation then is not only explanatory, it is the best way to explain such phenomena, by drawing out the characteristics of the system.

A further feature of the systems explanation is that to pick out a phenomenon as systemic is to adopt a certain kind of stance towards it. Dynamic systems theory offers an array of conceptual resources that can be put to explanatory use. Offering a systems explanation of some phenomenon involves a commitment to this being the kind of phenomenon that will be well-explained by the systems approach. In the introduction to this discussion, I mentioned in passing that some dynamicists, those who claim that natural cognitive systems are dynamical systems and are best understood from the perspective of dynamics, argue that labelling something a dynamical system is non-trivial and that it is an empirical question whether or not that thing is a dynamical system. The answer to that empirical question (which Port and Van Gelder thought was whether the system's future states are uniquely determined by its current states in accordance with some rule) will presumably be what justifies the adoption of the systemic stance towards the target phenomenon. The dynamicist's claim is controversial, and I don't intend to defend it, but merely to note that *if* calling on systems explanation involves adopting a kind of stance towards some phenomenon, then the adoption of that stance will presumably be justified by the answer to the "empirical question" that the dynamicist argues can be asked about whether or not the given activity forms a dynamical system. Use of "stance" language of course brings Dennett's position on the "intentional stance" to mind[28], but note that an endorsement of systemic stance-taking isn't an endorsement of Dennettian instrumentalism about systems. If there is an answer to this "empirical question" about what marks out a system then the status of systems will be far from instrumental.

I have also already discussed the idea that the system stance brings with it an expectation of reciprocal co-variation between variables. This is one candidate for the "empirical question",

[28] See Dennett 1987.

although, as I previously discussed, I don't expect reciprocity to be a completely reliable indicator of a system, but claim instead that understanding something as systemic brings with it an assumption about this relation of interactive reciprocity between variables.

Having laid out this positive account, I will now address two alternative accounts of the explanatory role of DST-C models. The first is a general account of models, developed by Morgan and Morrison, on which models act as mediators between theory and world. I will argue that DST-C models don't fall into this category. The second is a more specific claim about the HKB model – Clark argues that the HKB model *"owes its status as an explanation to its ability to illuminate what philosophers call counterfactuals"*[29]. I will argue that we should be wary of identifying predictive power with explanatory power, and that prediction alone does not qualify the HKB model as a scientific, or even a cognitive-scientific, explanation.

DST-C models NOT as mediators

> *"… we want to outline… an account of models as autonomous agents, and to show how they function as instruments of investigation. We believe there is a significant connection between the autonomy of models and their ability to function as instruments. It is precisely because models are partially independent of both theories and the world that they have this autonomous component and so can be used as instruments of exploration in both domains."*[30]

Morrison and Morgan defend an account of models whereby models act as mediators between theory and world. As is mentioned above, the key aspect of this relationship is the *autonomy* and *independence* of models from both theory and world. Models are independent in virtue of their construction, function, representational role and facilitation of learning. Take the case of construction. Here Morgan and Morrison argue that models are constructed out of a mixture of elements – descriptions from the empirical domain, mathematical representations, background theory – and that this mixed etiology renders them independent of both theory and

[29] Clark 1997a pg 117

[30] Morgan and Morrison 1999 pg 10

data. The independence in turn then allows them to mediate between theory and data, partly by acting as instruments, sometimes of exploration, sometimes of measurement and sometimes acting as "*technologies for investigation"*. Consider the following example:

> *It is possible using a plane pendulum to measure local gravitational acceleration to four significant figures of accuracy. This is done by beginning with an idealised pendulum model and adding corrections for the different forces acting on various parts of the real pendulum. Once all the corrections have been added, the pendulum model has become a reasonably good approximation to the real system… although we use the real pendulum to perform the measurement, that process is only possible given the corrections performed on the model. In that sense the model functions as an instrument that in turn enables us to use the pendulum to measure G*[31].

In this case, the model's independence facilitates its function as enabling a certain strand of measurements.

I will argue that DST-C models *don't* have the autonomous status of mediators in Morgan and Morrison's sense, and furthermore don't have the range of functions attributed to mediators. That is not to say that other models, including those discussed by Morgan and Morrison, don't have this range of functions and thus don't deserve the title "mediators". It should be noted, however, that if you accept the view that DST-C models don't act as mediators, then you can't take Morgan and Morrison's view that *all* models should be characterised as mediators. This discussion may end up bringing us to the more plausible view that the notion of *mediator* captures a range of features that many models have, but that few models have all of them.

There is a difficulty inherent in this discussion due to the singularly diverse nature of models in science. Morgan and Morrison are absolutely right that many scientific models are idealised, that many of them act as instruments and that, in virtue of their construction, many of them stand independent from data and theory, acting as mediators between the two. But consider the range of models they are working with – the MIT-Bag model of quark confinement, the model of pendulum motion discussed above, Fitzgerald's mechanical models of the aether, Marx's economic models – this is a random selection of the scientific models they mention. Their

[31] Ibid pg 22

account encompassed models made of wood and elastic bands and models made of equations. I would argue that they run into trouble by trying to get their account to cover simply too much stuff. Morgan and Morrison would, however, still argue that the DST-C models I have discussed are mediators, and I still have to give my justification for disagreeing with them.

I'll take as an example the model that Morgan and Morrison deal with that is closest in format to the HKB model of phase transitions in finger movements. This is the mathematical model of the business cycle, described during their discussion of a study of model construction:

> *In order to build a mathematical model of the business cycle, the economists that he studied typically began by bringing together some bits of theories, some bits of empirical evidence, a mathematical formalism and a metaphor which guided the way the model was conceived and put together. These disparate elements were integrated into a formal (mathematically expressed) system taken to provide the key relationships between a number of variables. The integration required not only the translation of the disparate elements into something of the same form (bits of mathematics) but also that they be fitted together in such a way that they could provide a solution equation that represents the path of the business cycle.*[32]

I take this to be close to the HKB model due to their overall similar aims – taking some systemic phenomenon and describing the relations between key variables in mathematical terms.

This description of the construction of the model is intended to reveal the roots of the autonomous status of scientific models. In virtue of its being constructed from such diverse elements as data, formalism and metaphor, the model stands independent from both theory and world and so can function as an autonomous mediator in Morgan and Morrison's sense, where this function is also supported by further features of the resulting model such as its representative role and its capacity as a tool for learning. The HKB model is similarly constructed – a given bit of behaviour is broken down into a series of interactions between variables, and patterns in those interactions are then captured in a mathematical formalism. I would argue, however, that in formulating the HKB model Kelso and his colleagues didn't put together something independent of both data and theory. The HKB model is simply a presentation of the basic data within the conceptual framework provided by dynamic systems theory, where the conceptual framework is

[32] Ibid pg 13

itself intended to capture features of the activity of systemic phenomena and stands and falls by its empirical success as such. Rather than seeing the conceptual framework as being imposed on the data ("top-down", if you like) we should understand the conceptual framework as arising out of previous observation of similar data ("bottom-up"), and the application of this framework to a given set of phenomena is simply an acknowledgement that here is some systemic behaviour of a kind that has been observed before. On this account, the DST-C model is nothing more than a description of the phenomenon in terms of a conceptual framework, where the framework itself has been developed partly in light of previous dealings with relevantly similar phenomena. Putting it in these terms makes the construction of the model seem less conceptually loaded than the Morgan and Morrison account would suggest, and it doesn't yield the required autonomy from theory and data that the status of "model as mediator" requires on their view. I would be hesitant to claim that *no* models are mediators in their sense, but it seems clear that DST-C models don't have the autonomy from theory and data required to support this status.

A further important difference between the HKB model and Morgan and Morrison's notion of models as mediators is the role of metaphor in the mediator, which is absent in the HKB model. In the example of the business cycle model, the construction of which is described above, the economists are described as constructing the model out of a mixture of elements, including a metaphor that dictates how the elements should be brought together. This is not the case in the HKB model; indeed, the HKB model has a notably un-mixed etiology. It captures patterns in the relations between activity in the relevant variables, but neither theory nor metaphor dictates the construction of the model. Rather, the theory comes in after the construction, in interpreting the significance of the model (for instance, in describing some state as an attractor). Indeed, metaphor can also come in at the interpretative stage (for instance, when Thelen compares mental activity

to *"a mountain stream flowing over a rocky bed*[33]*"*), but its place is in the interpretation, rather than the construction, of the model.

Predictive power and the "illumination" of counterfactuals

Clark discusses the HKB model as part of his discussion of the role of dynamic systems theory as a tool for what he calls "emergent explanation", one of the three kinds of explanation he argues are necessary for successful cognitive science. Having discussed the predictive capacity of the HKB model – it both reproduces the results of minor interferences in the system and the time taken to switch phases under different conditions[34] – Clark makes the following claim:

> *It should be clearer now why the dynamical account is not merely a nice description of the observed phenomena. It owes its status as an explanation to its ability to illuminate what philosophers call counterfactuals: to inform us not just about the actual observed behaviour of the system but also about how it will* (sic) *behave in various other circumstances*[35].

It is a mistake, however, to take the predictive power of the HKB model as sufficient for scientific explanation, or even cognitive scientific explanation. The following example illustrates my point.

Imagine that I have a deep curiosity about my neighbours' behaviour. Every morning I sit by the net-curtained window and document activity on the street. I take note of the time at which various people leave their houses and also take note of postal, milk and newspaper deliveries. I have done this for many years now and have an extensive catalogue of my neighbours' movements during that time. I have also noticed correlations between certain events, for instance, that nobody leaves the house for work at number 89 on every last Friday of the month, even though the woman who lives at number 89 leaves the house at 8.30 am every other weekday.

[33] Port and Van Gelder eds 1995 pg 71
[34] See Kelso 1995 pg 58-59 for details

[35] Clark 1997a pg 117

When it comes to the last Friday of the month, I can therefore successfully predict that the woman at number 89 will not leave the house at 8.30. This is clearly a case of prediction without explanation.

You could perhaps argue that this isn't fair; this isn't the interesting kind of prediction that Clark was talking about. It would be far more impressive if my almanac of street events could help me to predict the effect of *interferences* in the street "system". Consider the following, then: I have noticed that every time someone on the street goes on holiday, their newspaper deliveries stop. It is July, and, like every year, I have observed the family at number 96 pack their car with buckets, spades, kites and so on before heading off for their usual two weeks at the seaside. I can make a "boring" prediction that newspaper deliveries to number 96 will stop for the next two weeks. I can also, however, predict that *if* the family at number 96 were to come home from their holiday before the two week period was up, *then* the newspaper deliveries would re-start before the end of the two-week period. The family at number 96 have never come home from their holiday early, and let's say for the sake of argument that neither has anyone else I have observed on the street. The correlation between this family being in residence and their having newspapers delivered is high enough for me to make this prediction, to "illuminate the counterfactual", in Clark's terms. What I haven't, notably, done here is provided any explanation of the behaviour[36].

What this journey into armchair anthropology has shown is that we should be wary of identifying predictive power with explanatory power. The two will come hand in hand - indeed I would say that predictive power is necessary for explanation - but it is certainly not sufficient. It is important to note at this stage that Clark's comments about prediction came up in the context of a discussion of styles of explanation in cognitive science, rather than in a more general discussion of scientific explanation. We should expect a model of cognitive-scientific explanation to differ from a model of scientific explanation in significant and interesting ways, but I would argue that

[36] Traditional counterexamples to the D-N model of scientific explanation also give the result of prediction without explanation. I can, for instance, predict a storm from the barometer's falling but the barometer falling doesn't explain the storm.

predicting the result of interventions in some system is a very weak standard for explanation in *any* domain. The explanatory power of the DST-C model lies in its providing resources for a systems explanation, bringing with it a range of conceptual resources and an explanation of the phenomenon in terms of a conceptual backdrop that affords the adoption of a systemic stance towards it. The systems explanation will enable predictions, but the explanatory role of the HKB model isn't exhausted by its predictive power.

Conclusion

I started with the question: *How can a description of a thing ever act as an explanation of it?* I have attempted to answer this question for a single case, the HKB model of finger phase transitions, asking what explanatory role this model plays. I defended an account whereby the HKB model provides resources for systems explanations, where a systems explanation is a scientific explanation which proceeds by giving a profile of an interactive system that gives rise to the phenomenon that you are trying to explain. I argued that systems explanations should be understood as legitimate scientific explanations, as taking some phenomenon to be systemic brings with it an assumption that it will involve reciprocal co-variation between variables that can only be explained at a whole systems level.

In developing this account, I have left a number of questions open. What is a system? Why should we expect reciprocal interactions between the variables in a system? Is "system" a natural kind, or even an explanatory one? Or is "system" just a trivial tag? Answers to these questions will be of great relevance to ideas about the role systems explanations as I have described them. My aim, however, was to argue that, given the assumption that we do label phenomena as systemic, no matter what our justification for that labelling practice is, systems explanations should be viewed as legitimate scientific explanations, and furthermore, that the role of the HKB model of phase transitions in human finger movements is to provide resources for systems explanation.

Bibliography

1. Bechtel and Abrahamsen 2006. *Phenomena and Mechanisms*, available at http://mechanism.ucsd.edu/~bill/research/controversiespaper.pdf. Reprinted in Stainton, ed, *Contemporary Debates in Cognitive Science*, Oxford 2006.

2. Bechtel, Mandik, Mundale, Stufflebeam, eds. 2001. *Philosophy and the Neurosciences: A Reader*, Blackwell.

3. Clark 1997a. *Being There*, MIT.

4. Clark 1997b. The Dynamical Challenge, *Cognitive Science Vol 21 (4), 461-481.*

5. Cummins, 1975. Functional Analysis, *Journal of Philosophy Vol 72 (20), 741-765.*

6. Dennett, 1987. *The Intentional Stance*. MIT.

7. Eliasmith 2003. Moving Beyond Metaphors: Understanding the Mind for What it Is, *Journal of Philosophy 100 (10) 493-520.* (I used the draft available at http://arts.uwaterloo.ca/~celiasmi/Papers/eliasmith.moving%20beyond%20metaphors.jphil.pdf.)

8. Kelso, 1995. *Dynamic Patterns*: *The Self-Organization of Brain and Behavior*, MIT.

9. Morgan and Morrison, eds. 1999. *Models As Mediators*: *Perspectives on Natural and Social Science*, Cambridge.

10. Port and Van Gelder, eds. 1995. *Mind As Motion: Explorations in the Dynamics of Cognition*, MIT.

11. Salmon, 1989. *Four Decades of Scientific Explanation*, Pittsburgh.

12. Van Fraassen, 1980. *The Scientific Image*, Oxford.

13. Woodward, 2003. *Scientific Explanation*, Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/scientific-explanation/