The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective*

David Michael Kaplan and Carl F. Craver†‡

We argue that dynamical and mathematical models in systems and cognitive neuroscience explain (rather than redescribe) a phenomenon only if there is a plausible mapping between elements in the model and elements in the mechanism for the phenomenon. We demonstrate how this model-to-mechanism-mapping constraint, when satisfied, endows a model with explanatory force with respect to the phenomenon to be explained. Several paradigmatic models including the Haken-Kelso-Bunz model of bimanual coordination and the difference-of-Gaussians model of visual receptive fields are explored.

1. Introduction. In many areas of science, explanations are said to be adequate to the extent, and only to the extent, that they describe the causal mechanisms that maintain, produce, or underlie the phenomenon to be explained, the explanandum phenomenon. This form of explanation has received sustained attention in the philosophy of science (e.g., Bechtel and Richardson 1993/2010; Machamer, Darden, and Craver 2000; Craver

*Received August 2010; revised November 2010.

†To contact the authors, please write to: David Michael Kaplan, Department of Anatomy and Neurobiology, Philosophy-Neuroscience-Psychology Program, Washington University School of Medicine, Box 8108, 660 South Euclid Avenue, Saint Louis, MO 63110; e-mail: kaplan@eye-hand.wustl.edu. Carl F. Craver, Department of Philosophy, Philosophy-Neuroscience-Psychology Program, Washington University in St. Louis, 1 Brookings Drive, Wilson Hall, Saint Louis, MO 63130.

‡Thanks to Lindley Darden, Peter French, Peter Machamer, Gualtiero Piccinini, Dan Weiskopf, and two anonymous referees for helpful comments on an earlier draft of this article.

Philosophy of Science, 78 (October 2011) pp. 601–627. 0031-8248/2011/7804-0004$10.00
Copyright 2011 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

2007) and has fruitfully been applied to exemplars of explanations in molecular biology and genetics (Darden and Craver 2002; Darden 2006), electrophysiology (Craver 2006, 2007), evolutionary biology (Baker 2005; Skipper and Millstein 2005), many areas of social science (Hedström and Ylikoski 2010), and empirically driven economics (Craver and Alexandrova 2008). It is an open question whether this philosophical view of explanation also expresses desirable norms of explanation for dynamical and mathematical models in the domains of systems neuroscience and cognitive neuroscience and, if so, precisely what one demands by insisting that explanations in such sciences must describe mechanisms.

Here, we articulate and defend a mechanistic approach to explanation for dynamical and mathematical models in systems neuroscience and cognitive neuroscience. Such models, like models in "lower-level" neuroscience, carry explanatory force to the extent, and only to the extent, that they reveal (however dimly) aspects of the causal structure of a mechanism. This view contrasts with dynamicist views, according to which explanations need not respect the underlying causal structures that give rise to system-level dynamics (e.g., van Gelder 1995, 1998; Chemero and Silberstein 2008). We argue that there is no currently available and philosophically tenable sense of 'explanation' according to which such models explain even when they fail to reveal the causal structures that produce, underlie, or maintain the explanandum phenomenon. Dynamicists' models at their best are descriptive tools for representing how complex mechanisms work; they are part of, not an alternative to, the project of providing mechanistic explanations. Those who defend the dynamical approach as a successor to outmoded, mechanistic explanation fundamentally misidentify the source of explanatory power in their models.

We begin with a reminder from the past 6 decades of philosophical work on scientific explanation (e.g., Salmon 1989): not all empirically (descriptively or predictively) adequate models explain. Furthermore, the line that demarcates explanations from merely empirically adequate models seems to correspond to whether the model describes the relevant causal structures that produce, underlie, or maintain the explanandum phenomenon. This demarcation line is especially significant as it also corresponds to whether the model in question reveals (however dimly) knobs and levers that potentially afford control over whether and precisely how the phenomenon manifests itself (Woodward 2003). We summarize these considerations in a model-to-mechanism-mapping (3M) requirement, accordingly. The 3M constraint is the mechanist's gauntlet: a default assumption that the phenomena of cognitive and systems neuroscience have mechanistic explanations, like so many other phenomena in the special sciences, and that cognitive and systems neuroscientists ought to (and often do) demand that explanations reveal the mechanisms underlying the phenom-

https://doi.org/10.1086/661755 Published online by Cambridge University Press

ena they seek to explain. Like all default stances, 3M is defeasible. However, those who would defease it must articulate why mechanistic styles of explanation are inappropriate, what nonmechanistic form of explanation is to replace it, and the standards by which such explanations are to be judged.

While our view is conservative, it is not imperialistic. Specifically, we do not intend 3M to rule out nonmechanistic explanation generally. There might be domains of science in which mechanistic explanation is inappropriate.[1] Justified in part by the stunning success of the mechanistic tradition in the lower-level domains of neuroscience and in the special sciences more generally, 3M is rather a specific recommendation for thinking about the explanatory aspirations of cognitive and systems neuroscience.

2. Nonmechanistic Explanation: The Very Idea. Our positive thesis is clearly opposed to those who hold, in whatever terms, that mechanistic explanation is no longer an appropriate goal for cognitive and systems neuroscience. In particular, we oppose strong dynamicist and functionalist views according to which mathematical and computational models can explain a phenomenon without embracing commitments about the causal mechanisms that produce, underlie, or maintain it.[2]

Chemero and Silberstein (2008), for example, argue that while mechanistic explanation is appropriate in lower-level neuroscience (i.e., for such phenomena as action-potential propagation, gene regulation, and neurotransmitter release), it is inappropriate for explaining higher-level dynamical or cognitive systems. They state: "A growing minority of cognitive scientists, however, have eschewed mechanical explanations and embraced dynamical systems theory. That is, they have adopted the mathematical methods of nonlinear dynamical systems theory, thus employing differential equations as their primary explanatory tool" (11). In advocating dynamic systems theory as an alternative to prevailing views of explanation, their view falls neatly into a long tradition of noncomputationalist and noncognitivist approaches to the explanation of cognition (e.g., van

1. It is often said that certain areas of physics require explanations that do not involve decomposing phenomena into component parts (see Bechtel and Richardson 1993/2010; Glennan 1996). Others hold that mental phenomena, such as belief and inference, are fundamentally normative and so demand noncausal forms of explanation (McDowell 1996).

2. We focus only on dynamicist opponents of the mechanistic approach to explanation. Functionalists have similarly argued that functional analysis encompasses a nonmechanistic explanatory paradigm or framework in its own right. For further discussion of mechanistic explanation in neuroscience vis-à-vis functional explanation, see Piccinini and Craver (forthcoming).

https://doi.org/10.1086/661755 Published online by Cambridge University Press

Gelder 1995, 1998; Haugeland 1998b).[3] It is undeniable that dynamical models of brain function and cognition, in which the temporal evolution of system-level variables is mathematically described using differential equations, are now commonplace in systems and cognitive neuroscience (e.g., Schöner and Kelso 1988; Jirsa, Fuchs, and Kelso 1998; Kelso et al. 1998; Fuchs, Jirsa, and Kelso 2000a, 2000b; Bressler and Kelso 2001; Carson and Kelso 2004; Oullier et al. 2008; Jantzen, Steinberg, and Kelso 2009; Tognoli and Kelso 2009; Kelso 2010). However, the fact that this practice is entrenched, or even that it plays an increasingly indispensable role in contemporary science, does not establish that models of this sort have explanatory force independent of whether they describe mechanisms as Chemero and Silberstein maintain. To justify this further claim, it is necessary to articulate a nonmechanistic form of explanation and, more importantly, to clarify the rules by which such explanations will be judged.

A central thrust in this line of argument has been that dynamical models characterize the behavior of systems, not in terms of their component parts but in terms of emergent or higher-level variables describing global states of the system. Thelen and Smith (1994) argue that such global behavioral patterns in complex systems obtain "irrespective of their material substrates." Such general patterns can be found in "systems of different levels of diversity and complexity, and whose constituent elements are completely dissimilar" (49). Chemero and Silberstein echo this view in claiming that the "key feature of such dynamical explanatory models is that they allow one to abstract away from causal mechanical and aggregate micro-details to predict the qualitative behavior of a class of similar systems" (2008, 12). If these dynamicists are right, such models yield explanations in the total absence of commitments regarding the causal mechanisms that produce the cognitive or system behavior we seek to explain.

If the explanatory force of these models does not arise from the fact that they describe mechanisms, as the old tradition would have it, then what is the source of their explanatory power? Some dynamicists seem to hold that the explanatory power of these models follows from their descriptive and predictive power. Port and van Gelder (1995) stress that

3. Dynamic systems theory is a branch of mathematics studying the general properties of dynamic systems such as how the state of a complex system evolves over time. Beyond using differential or difference equations to describe the temporal evolution of a system's state, dynamic systems theory also possesses geometric tools for representing patterns of change in the state of a system over time in terms of trajectories through phase or state space, where each point in this abstract space defines a possible state of the system. Although a system can in principle take any path through its state space, it may also be driven toward an attractor, a set of points toward which a dynamical system evolves over time. For further discussion, see van Gelder and Port (1995).

https://doi.org/10.1086/661755 Published online by Cambridge University Press

dynamical explanation "yields not only precise *descriptions* . . . but also *predictions* which can be used in evaluating the model" (15). In a seminal article that has become a manifesto for the dynamical approach in cognitive science, van Gelder asserts that "many factors are relevant to the goodness of a dynamical explanation, but the account should at least capture succinctly the relations of dependency, and make testable predictions" (1998, 625). Chemero and Silberstein also connect explanation and prediction: "If models are accurate enough to describe observed phenomena and to predict what would have happened had circumstances been different, they are sufficient as explanations" (2008, 12). Similarly, Walmsley (2008) argues that dynamical explanations in cognitive science are covering law explanations and that dynamical models should be regarded as conforming precisely to the conditions on adequate explanations laid out by the covering law model (Hempel 1965; see also Bechtel and Abrahamsen 2002). According to the covering law model, explanations are structurally equivalent to predictions; each involves showing the phenomenon to be an instance of a regular, repeated pattern. Dynamicists thus emphasize that their models have predictive scope across superficially distinct systems; the wide predictive scope of the model, its application to many different systems of a similar global type, is the source of their explanatory power. These dynamicists appear to be predictivists about explanation.

3. Mechanistic Explanation in Neuroscience and the 3M Constraint. In contrast to predictivism, mechanists insist that models have explanatory force in virtue of the fact that they describe the causes and mechanisms that maintain, produce, or underlie the phenomena in a given domain. From this perspective, a model can save the phenomenon tolerably well and yet fail to explain how the system behaves. To explain the phenomenon, the model must in addition reveal the causal structure of the mechanism. This will involve describing the underlying component parts, their relevant properties and activities, and how they are organized together causally, spatially, temporally, and hierarchically. This view of mechanistic explanation has been developed at length elsewhere (see Bechtel and Richardson 1993/2010; Machamer et al. 2000; Craver 2007; Bechtel 2008). We here emphasize only that the term 'mechanism', as developed by this tradition, has been liberated from many of its misleading historical associations. First, to insist on mechanistic explanations is not to insist on explanation in terms of simple machines governed by strict deterministic laws or in terms of physical contact, energy conservation, or any other fundamental or otherwise privileged set of activities. Second, mechanistic explanation is not only downward looking, peering down into the mechanisms within mechanisms by which things work, but it is also contextual,

https://doi.org/10.1086/661755 Published online by Cambridge University Press

situating mechanisms within higher-level causal structures. The parts need not be spatially localized within the system. Nor need their activities be sequential, from beginning to end; they might involve (negative or positive) feedback loops or recurrent connections between components. Frequently, features of the spatial and temporal or dynamic organization of the components and their activities are explanatorily relevant and so are included in the models.[4] Other times, it matters more who communicates with whom than precisely where the participants are located (to borrow a metaphor from Haugeland's [1998a] insightful discussion of this form of explanation). Finally, mechanisms are frequently described using equations that represent how the values of component variables change with one another. Mathematical description, while not essential to all mechanistic explanations, is certainly a useful tool for characterizing the complex interactions among components in even moderately complicated mechanisms. If a phenomenon is characterized as an input-output relationship to be explained, mechanistic explanations describe the relevant causes lying between the input and the output.

What remains of mechanistic explanation from the days of Descartes and Boyle is this: that one explains a phenomenon by showing how it is situated in the causal structure of the world (Salmon 1984). In the downward-looking aspect with which we are most concerned, one reveals the internal causal structure of a phenomenon, explaining the features of the phenomenon in terms of the activities of, and causal relations among, the component parts.

The primary virtue of the causal or mechanistic view of explanation, and one reason why it is the dominant view of explanation in the philosophy of science at present, is that it neatly dispenses with several well-known problems of predictivism: the thesis that the explanatory force of a model derives from its predictive power. Judging from the above passages, many dynamicists are predictivists. Predictivists of the most extreme form hold that any predictively adequate theory is, ipso facto, explanatory. A predictivist might hold that to explain a phenomenon is to show that its description follows from universal or statistical generalizations conjoined with descriptions of the initial and boundary conditions (Hempel 1965; for further discussion, see Douglas 2009). They might hold that any predictively adequate model (i.e., any model that predicts the relevant aspects of the phenomenon with the required precision and accuracy) explains those aspects of the phenomenon. For the predictivist, explanation is not an additional virtue beyond saving the phenomena; the phenomena are explained as soon as they are saved.

4. See Bechtel and Abrahamsen (2010) for examples of this kind.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

Predictivism, in its extreme form, runs counter to the scientific-commonsense judgment that only some models are explanations. Common sense is often misleading, and our question cannot be settled with a poll. However, the scientific-commonsense judgment returned in this case hits on a set of distinctions that have been crucial to the advance of science generally and of the special sciences in particular. In reviewing these commonsense judgments, we are sketching the contours of a principled and pragmatically significant division in the space of empirically (and predictively) adequate models: between those that describe the causes that produce, underlie, or maintain the phenomena and those that do not.

Many of the well-known counterexamples to the covering law model of explanation (Salmon 1989) attack predictivism directly. One can predict a storm from the falling mercury in a barometer, but the falling mercury does not explain the storm. One can predict a nearly empty gas tank from the sputtering of the engine, but the sputtering does not explain why the tank is nearly empty. One can predict that the ball game will begin from the performance of the national anthem, but the performance of the national anthem does not explain the start of the game. These commonsense judgments seem to rely on the underlying premise that explanations correctly identify features of the causal structures that produce, underlie, or maintain the explanandum phenomena. Predictivists must either deny these commonsense judgments or offer an alternative view of explanation that yields those judgments as well. Dynamicists, as predictivists, also face this choice.

Consider a second apparent problem for predictivism, also grounded in Hempel's pioneering discussion of the covering law model. The predictive value of a model is unaffected (at least in many cases) by the inclusion of irrelevant detail. If one set of law statements and boundary conditions, K, entails another set, P, then the conjunction $K \wedge S$, where S is any sentence that fails to contradict a member of K, also entails P. Hempel called this the problem of irrelevant conjunction (1965, 273 n. 33). It poses a problem because it conflicts with the common scientific practice of filtering out irrelevant factors from our explanation. An explanatory model suffers, for example, if it includes irrelevant parts that are not in the mechanism, irrelevant properties that play no causal role, or irrelevant activities that are sterile in the mechanism. The norms for evaluating explanatory models are more strict than the norms for evaluating predictive models. Thus, explanation is not (merely) prediction.

Let us now come at the issue from a third perspective. Scientists commonly draw a distinction between models that merely describe a phenomenon and models that explain it. Neuroscientists such as Dayan and Abbott, for example, distinguish between purely descriptive mathematical models, models that "summarize data compactly," and mechanistic mod-

https://doi.org/10.1086/661755 Published online by Cambridge University Press

els, models that "address the question of how nervous systems operate on the basis of known anatomy, physiology, and circuitry" (2001, xiii). Mechanistic models describe the relevant causes and mechanisms in the system under study. In their view, a predictively adequate mathematical model might fail to characterize the relevant neural mechanisms.

The distinction between purely descriptive, phenomenal models and mechanistic models is familiar in many sciences. Snell's law describes how light refracts as it passes from one medium to another, but the law does not explain why the path of light changes as it does. It merely expresses the regularity in need of explanation. To explain this principle, one must appeal to facts about how light propagates or about the nature of electromagnetic phenomena.[5] The Boyle-Charles model of ideal gases describes a mathematical relation between pressure, volume, and temperature. The kinetic theory of gases, in contrast, posits a mechanism to explain why the gas law holds. This distinction has also played a crucial role in lower-level neuroscience. Hodgkin and Huxley (1952) generated equations to model how the conductance of a neuronal membrane to sodium and potassium changes as a function of voltage during an action potential, but they did not explain how voltage changes membrane conductance. The explanation required the idea of a voltage-sensitive, membrane-spanning channel, which only came dimly into view in the 1970s and 1980s (Hille 1992; Bogen 2005, 2008; Craver 2006, 2007, 2008). Bliss and Lomo (1973) described long-term potentiation qualitatively, allowing them to predict that hippocampal synapses will be strengthened as a result of rapid and repeated stimulation, but they could not yet explain why synapses exhibit this phenomenon. The signature of a phenomenal model is that it describes the behavior of the target system without describing how the mechanism underlying that behavior works. Predictivists thus have difficulty recognizing the distinct explanatory value assigned to models that go beyond providing a generalized description of a phenomenon (such as Hodgkin and Huxley's conductance equations) and those that reveal the causal structures responsible for the phenomenon.

Two related points follow this line of thought. First, the strong predictivist has difficulty expressing the explanatory limits of mere how-

5. Of course, one might explain the angle of refraction of a given beam of light by appeal to the fact that the light crossed between two media in which it has different velocities. However, we are interested here in explaining why light generally bends when it passes from one medium to the next. Likewise, one might explain the expansion of a balloon by reference to the increase in temperature and the relevant causal counterfactual that gases expand when heated, but we are here interested in explaining why gases expand when heated. It is not explanatory to tell us that all ideal gases do so. Such a response would merely inform us about the scope of our request for explanation.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

possibly models or theories (Dray 1957).[6] Mere how-possibly models describe mechanisms that could produce the phenomenon in question but that, in fact, do not produce the phenomenon. The human visual system could be constructed out of photovoltaic cells and Tinkertoys, yet that is not in fact how it works. Scientists often construct how-possibly models as a strategy of surveying the space of possible mechanisms capable of producing a phenomenon.[7] Building such models can provide further mechanistic hypotheses to test, as well as clues about general principles governing the organization of a given mechanism (or class of mechanisms). In scientific common sense, however, there is a vast gulf between providing a model that might explain a phenomenon and explaining the phenomenon. For the strong predictivist, however, if two models account for the phenomenon equally well, then they would be explanatorily equivalent. Repeated calls for biological realism in modeling can be seen as expressing the opposed judgment that not all empirically adequate how-possibly models are equally explanatory. This judgment insists on a mapping between items in the model and components, properties, and relations in the system under study.

Second, commonsense judgment allows that one might increase the quality of one's explanation without improving the predictive reach of one's models. For example, different models of a mechanism that are equally empirically adequate might be more or less complete. One model might be more of a mechanism sketch, identifying one or two significant internal variables from which the vast majority of the variance in the phenomenon can be accounted, without understanding the causal structures by which those variables change or by which those variables influence the phenomenon. Hodgkin and Huxley posited "activation particles" to explain how membranes change their conductances, but they knew that this was only a filler term for a mechanism-we-know-not-what (Hodgkin and Huxley 1952; Hodgkin 1992). As models of channel physiology began to appear in the 1980s, researchers were making explanatory progress by revealing internal parts, properties, and organizational features, without corresponding gains in predictive adequacy over the Hodgkin-Huxley model (Hille 1992; Doyle et al. 1998).

These last two points require more qualification than we can reasonably offer in this article. For example, the idea of an ideally complete how-actually model, one that includes all of the relevant causes and compo-

6. We thank Alex Rosenberg for bringing Dray's (1957) discussion of how-possibly explanations to our attention.

7. For example, consider Bertil Hille's (1992) many diagrams of possible mechanisms for the sodium channel and Watson and Crick's failed spatial models of DNA (discussed in Darden 2006).

https://doi.org/10.1086/661755 Published online by Cambridge University Press

nents in a given mechanism, no matter how remote, negligible, or tiny, without abstraction or idealization, is a philosopher's fiction. Science would be strikingly inefficient and useless both for human understanding and for practical application if it dealt in such painstaking minutiae. The striking achievements of science, such as Hodgkin and Huxley's model of the action potential, require that one make idealizing assumptions (e.g., that the axon is a perfect cylinder) and that one gloss over minor details to capture broad and robust patterns in the causal structure of a mechanism. In the special sciences generally, and cognitive and systems neuroscience particularly, one should expect considerable variability in the mechanisms for a given phenomenon across individuals and in the same individual over time. The special sciences would be utterly paralyzed if complete how-actually explanations were the guiding objective. Yet these commonplace facts about the structure of science should not lead one to dispense with the idea that models can more or less accurately represent features of the mechanism in the case at hand and that models that describe more of the relevant features of the mechanism are more complete than those that omit them. These practices of abstraction and idealization sit comfortably with the realist objectives of a mechanistic science.[8]

Let us then return to the central point of this section, which is to reiterate some of the central shortcomings of predictivism. These shortcomings explain why causal and mechanistic views of explanation have become increasingly popular among philosophers of science. If one thinks that explanations describe causes and mechanisms, then one can rule out, as nonexplanatory, models that explain one variable by appeal to another variable that is merely correlated with it and that propose to explain effects in terms of their causes. One can allay the problem of irrelevant conjunctions by allowing causal and constitutive relevance to stand as an indicator of explanatory relevance (see Salmon 1984; Craver 2007). One can distinguish phenomenal and explanatory models by insisting that the latter describe mechanisms. One can distinguish how-possibly models from models that describe the mechanisms that in fact maintain, produce, or underlie the phenomenon in question. And one captures a clear sense of explanatory completeness. We belabor these points, many of which have long been familiar to philosophers of science, because we think they are not given sufficient weight by those who recommend the abandonment of the mechanistic framework in neuroscience in favor of some other kind of explanation, such as dynamical explanations or functionalist expla-

8. Strevens (2004) discusses how to combine both causal-mechanical and unificationist approaches to explanation in a way that attempts to solve this problem. As it is beyond the scope of our current concerns, we do not, however, take up his proposed solution in any further detail here.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

nations. The ideal of mechanistic explanation, including the refusal to rest content with phenomenal descriptions, how-possibly models, and sketches, has guided the fundamental achievements made in low-level fields of neuroscience such as electrophysiology and molecular biology. Furthermore, advances in mechanistic explanation have revealed new knobs and levers in the brain that can be used for the purposes of manipulating how it and its parts behave. That is just what mechanistic explanations do.

Cognitive and systems neuroscientists have yet to make the kinds of inroads in their domain that Hodgkin, Huxley, Hille, and others made in understanding the electrical properties of neuronal membranes. Given that one expects cognitive mechanisms ultimately to be composed of lower-level mechanisms of this sort, in a manner that might be illustrated in a telescoping hierarchy of mechanisms and their components, it would be most tidy and parsimonious if the ideal of mechanistic explanation were to be extended from top to bottom across all fields in neuroscience. Our intended opponent suggests that cognitive and systems neuroscientists should abandon this set of explanatory ideals in favor of an alternative form of explanation. This suggestion, if it is to have content, should come with principled reasons for abandoning those ideals of explanation and crucially, an alternative vision of what the new rules of the explanation game are. And predictivism won't cut it.[9]

We summarize the above considerations in the form of a model-to-mechanism-mapping requirement that makes the commitment to a mechanistic view of explanation explicit:

> **(3M)** In successful explanatory models in cognitive and systems neuroscience (*a*) the variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (*b*) the (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism.

This principle is restricted to cognitive and systems neuroscience and so allows that there are legitimate nonmechanistic forms of explanation.[10]

9. The predictivist might require that explanations appeal to laws of nature (Rosenberg 2001) or perhaps that they unify diverse phenomena within a single explanatory framework (Kitcher 1989). Mechanists think such demands are ill fit to the explanatory domains of physiological and biological sciences (Bechtel and Abrahamsen 2005), and they doubt that such corrections can be made to predictivism without ultimately adopting some form of the mechanistic framework (Craver 2007).

10. Although providing a full defense for the claim is beyond the scope of this article, we see no good reason to exempt all of cognitive science from the explanatory demands laid out by 3M. To defend this claim adequately, we would have to answer the following sort of objection, raised by an anonymous referee: 3M is applicable to sciences, such

https://doi.org/10.1086/661755 Published online by Cambridge University Press

However, it suggests that cognitive neuroscience is a domain in which the demands of mechanistic explanation ought to be met. Second, the claim should be read as allowing for idealization and abstraction in models. As we mention above, the vagueness of the line between how-possibly and how-actually does not mean that there is no distinction to be had. Finally, we emphasize that models of mechanisms frequently involve mathematical descriptions of the causal relationships in a mechanism. This gives the misleading impression that explanation in such cases involves subsumption under generalizations and that it is the generalization that is doing the explanatory work. In fact, however, the generalizations are explanatory because they describe the causal relationships that produce, underlie, or maintain the explanandum phenomenon.

The commitment to mechanistic realism embodied in 3M is justified on several grounds. It is justified in part because it makes sense of scientific-commonsense judgments about the norms of explanation. It is also

---

as neuroscience, already committed to the paradigm of mechanistic explanation, but other sciences, such as cognitive science, recognize the possibility of explanations insensitive to the details about underlying implementation. Did not David Marr, after all, show us that it is possible to offer computational explanations at an algorithmic level, allowing all the while that the algorithm might be implemented in any number of disparate hardwares? And if so, is it not possible to explain without knowing the details of the mechanism? We certainly do not intend to deny the multiple realizability of the algorithms by which cognitive functions are performed. Given the variability among human brains, and how they change at the cellular and molecular levels from moment to moment, cognitive scientists and neuroscientists alike must describe neural mechanisms in a way that glosses over such fine-grained differences in implementation. That said, there remains a question of whether a given algorithmic description is the correct one for a given phenomenon. How is this question to be answered? One might claim that the correct algorithmic description is simply one that entails the various features of the explanandum phenomenon (i.e., one that can carry out the computation one seeks to explain). But as Marr also emphasizes, the same computation might be performed by any number of algorithms. And Marr would be the last person to suggest that any computationally adequate algorithm is equally good as an explanation of how a computation is performed in a given organism or species. Marr, it should be remembered, builds his vision algorithms on the basis of careful consideration of the properties of the early visual system such as the response profiles of retinal ganglion cells. Such a response, in other words, is simply a retreat to predictivism, and we have already explained why we think predictivism fails as a view of explanation. The other obvious alternative is that one should be able to identify features of the mechanism in the case at hand that can plausibly be interpreted as implementing a given algorithm: that the inputs are of the sort that the algorithm uses, that the outputs are of the sort that the algorithm posits, and that the transformation of one to the other is, in fact, carried out in the manner that the algorithm specifies. In short, we see no reason to exempt cognitive science from the explanatory demands of mechanism. It has not escaped our attention that 3M, should it be found acceptable, has dire implications for functionalist theories of cognition that are not, ultimately, beholden to details about implementing mechanisms. We count this as significant progress in thinking about the explanatory aspirations of cognitive science.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

justified by reference to the fact that an explanation containing more relevant detail about the components, properties, and activities that compose a given mechanism is more likely, all things equal, to be able to answer more questions about how it will behave in a variety of circumstances than is a model that does not aim at satisfying something like 3M. This follows from the fact that such models allow one to predict how the system will behave if parts are broken, changed, or rearranged and so how the mechanism is likely to behave if it is put in conditions that make a difference to the parts, their properties, or their organization.[11] The lesson is that if one values predictive power, one should seek mechanistic explanations that conform to 3M. Finally, models that conform to 3M reveal knobs and levers in mechanisms that might be used for the purposes of bringing the mechanism under our control (Woodward 2003). This connection between explanation and control might help to explain why scientific explanation remains prized as a distinct scientific virtue. Finally, the special sciences, and the physiological sciences in particular, have made steady advances over the past 2 centuries in providing explanations that approximate the ideals expressed in 3M. The mechanistic tradition should not be discarded lightly. After all, one of the grand achievements in the history of science has been to recognize that the diverse phenomena of our world yield to mechanistic explanation.

4. The Haken-Keslo-Bunz Model of Bimanual Coordination. To see how commitment to 3M might play out in cognitive and systems neuroscience, consider the HKB model (Haken, Kelso, and Bunz 1985) of the dynamics involved in human bimanual coordination. The HKB model has been the focus of extensive investigation in computational and systems neuroscience for over 20 years. Chemero and Silberstein (2008) cite the HKB model as evidence that some explanations in cognitive science and neuroscience are nonmechanistic and that for many complex behavioral and neural systems the "primary explanatory tools" are the mathematical methods of nonlinear dynamic systems theory.

The HKB model (eq. [1]) accounts for behavioral data collected when experimental subjects are instructed to repeatedly move their index fingers side to side in the transverse plane in time with a pacing metronome either in phase (simultaneous movements toward the midline of the body) or antiphase (simultaneous movements to the left or right of the body mid-

11. It is always possible (although never easy) to contrive a phenomenally adequate model post hoc if and when the complete input-output behavior of a system is known. However, the critical question is how readily we can discover this input-output mapping across the full range of input conditions without knowing anything about the underlying mechanism. We are far more likely to build predictively adequate models when aspects of the mechanism are known.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

line).[12] The metronome speed is an independent variable that researchers can systematically increase or decrease. When increased beyond a certain critical frequency, subjects can no longer maintain the antiphase movement and switch involuntarily into in-phase movement. Subjects who begin in phase do not shift. Only in-phase movement is possible beyond the critical frequency.

Dynamic systems theorists characterize this system in terms of state spaces and attractors and represent the behavior of the two moving fingers as coupled oscillators. At slow tempos, the oscillators can be stably coupled to one another in both antiphase and in-phase movement modes. The state space of the system is thus said to have two basins of attraction (or attractors)—one for antiphase and one for in-phase movement. At high tempos, in contrast, only the in-phase mode is possible—the state space has a single attractor. At the switch point, the attractor landscape of the system changes.

The HKB model provides an elegant mathematical description of the regularities that constitute this phenomenon, including the rate of change in the phase relationship between the left and the right index fingers and the critical switch point from the antiphase to the in-phase pattern. The core of the model is the differential equation describing the coordination dynamics of these coupled elements:

$$\dot{\phi} = -a \sin \phi - 2b \sin 2\phi, \tag{1}$$

where ϕ is the so-called collective variable representing the phase relationship (relative phase) between the two moving index fingers (when $\phi = 0$, the fingers are moving perfectly in phase), a and b are coupling parameters reflecting the experimentally observed finger oscillation frequencies, and the coupling ratio b/a is a control parameter since relatively small changes in its value can have a large impact on system behavior (for further discussion of the HKB model, see Haken et al. 1985; Kelso 1995; Bressler and Kelso 2001).[13]

12. In this literature, bimanual movements are commonly described in terms of the phase relation between left and right body parts (e.g., arms, hands, or fingers). A 0° phase difference between two body parts is conventionally defined as the state in which they are moving in a mirror-symmetric way in extrinsic space with respect the midline of the body. Alternatively, phase difference can also be specified according to a muscle-based definition of symmetry. In this case, a 0° phase difference would correspond to the simultaneous activation of homologous muscles in both effectors (Haken et al. 1985).

13. The coupling ratio b/a is inversely proportional to the movement rate or oscillation frequency. When its value is high, corresponding to a low movement rate, there are two stable attractors for the system dynamics. When b/a is low, only one stable attractor exists. At the critical value of b/a, the system undergoes an abrupt phase transition or bifurcation.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

Although, to our knowledge, Kelso and colleagues never explicitly discuss the explanatory value of their model, it clearly was not originally advanced as a description of the neural or biomechanical components responsible for the experimentally observed behavioral dynamics. As Bressler and Kelso (2001) describe it, the HKB model "exemplifies a law of coordination that has been found to be independent of the specifics of system structure. [It] captures the coordination between behaving components of the same system, between an individual and the environment, and even between two individuals socially interacting" (28). Kelso (1995) embraces a similar perspective: "I have to admit that one of the main motivations behind these experiments was to counter the then dominant notion of motor programs, which tries to explain switching (an abrupt shift in spatiotemporal order) by a device or mechanism that contains 'switches'" (57). Appearances notwithstanding, Kelso did not intend to replace one mechanistic hypothesis framed in terms of "switching mechanisms" with another. Kelso railed against all models of motor behavior that invoke underlying neural or computational processes such as the execution of underlying motor programs (i.e., the stored sequence of instructions or commands to drive the muscles during a movement): "any time we posit an entity such as reference level or program and endow it with content, we mortgage scientific understanding" (33–34). For Kelso, a predictivist on our spectrum, the HKB model was a full-fledged explanation, even though it does not describe mechanisms.

Kelso is right in one respect yet wrong in another. For reasons detailed below, he is right because loosely constrained conjectures about "switches" and "programs" are fictions, like activation particles in the Hodgkin-Huxley model. Such conjectures are often undischarged filler terms that may or may not correspond to components in the mechanism. However, he is incorrect because the differential equation for ϕ (eq. [1]) does not explain the phase transition. Despite the indisputable utility of dynamical modeling techniques as tools for describing phenomena, it is insufficient to explain a phenomenon merely to describe it in concise and general form. The HKB model is a mathematically compact description of the temporal evolution of a purely behavioral dependent variable (relative phase) as a function of another purely behavioral independent variable or order parameter (finger oscillation frequency). However, none of the variables or parameters of the HKB model correspond to components and operations in the mechanism, and none of the mathematical relations or dependencies between variables map onto causal interactions between those system components (as required by 3M). Variables in models of behavioral dynamics, such as HKB, might be said to involve macroscopic or behavioral "components," such as the phase relationship between the

https://doi.org/10.1086/661755 Published online by Cambridge University Press

fingers, but these are not components in the sense of being the underlying parts of the mechanism.

Van Gelder makes a similar point in his general discussion of dynamical models of cognition: "the variables [dynamical models] posit are not low level (e.g., neural firing rates) but, rather, macroscopic quantities at roughly the level of the cognitive performance itself" (1998, 619). If so, they are phenomenal models. They describe the phenomenon. They do not explain it any more than Snell's law explains refraction or the Boyle-Charles gas law explains why heat causes gases to expand. Accordingly, the HKB model does not reveal internal aspects of the system to take into account when making predictions, and it does not reveal new loci for intervening into the system to change the temporal location of the phase shift. It does not, therefore, deserve the honorific title 'explanation'.[14]

Van Gelder (1998) considers this objection explicitly. He acknowledges that mathematical equations merely constructed to fit a line to a set of data points provide at best a description rather than a genuine explanation. As he puts it, "A poor dynamical account may amount to little more than ad hoc 'curve fitting,' and would indeed count as mere description" (625). But a satisfactory dynamical account, he argues, should be no more suspect than many of our paradigms of scientific explanation: "Dynamical theories of cognitive processes are deeply akin to dynamical accounts of other natural phenomena such as celestial motion. Those theories constitute paradigm examples of scientific explanation. Consequently, there is no reason to regard dynamical accounts of cognition as somehow explanatorily defective" (625).

This response is inadequate at many levels. First, the history of astronomy is replete with models of celestial motion intended to save the phenomena without explaining them; think of Ptolemy's models or Tycho Brahe's ornate construction or the spirit with which Osiander introduced Copernicus's masterwork. The very distinction between saving the phenomena and explaining them traces to the history of such models. If dynamical models are like Ptolemy's, then they are empirically adequate

14. The equations describe how the shift from in to out of phase depends on the rate of oscillation, but they do not explain why the dependency holds. As noted above, one can explain why a subject's fingers go through the phase shift, by appeal to the increase in tempo and a causal generalization expressing the idea that the tempo makes a difference to whether the fingers are in phase or out of phase. We are concerned with the explanation of the relationship between tempo and phase, not with why a particular subject goes through the phase shift. This is just to restate the familiar distinction drawn by Salmon (1984) between etiological and constitutive aspects of causal-mechanical explanations. In this article, we are concerned only with the constitutive aspects.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

but explanatorily deficient. Second, van Gelder does not tell us how dynamical models and physical models are "akin" to one another. If the extent of their kinship is that they "make testable predictions" (1998, 625), are empirically adequate, or are succinct, then the arguments of section 3 show that this is insufficient to capture the norms for sorting good explanations from bad.

Walmsley (2008) follows van Gelder on this point, arguing that the HKB model explains to the extent it is capable of generating accurate quantitative predictions for the observed phase transitions in subject behavior, as well as generating predictions about unobserved effects on the behavioral dynamics induced by other experimental perturbations, subsequently confirmed in further experiments. Van Gelder suggests that an adequate model should describe "the relations of dependency," and one has the sense that he means by this that it should describe relations of counterfactual dependency.[15] If these statements are meant to suggest that the model should describe mechanisms, including relations of dependency among components, their properties, and their activities, then we agree that such models have explanatory force. If it is meant, however, to suggest that the model merely displays the dependency that defines the explanandum phenomenon, then we disagree that such a model explains the phenomenon. In our example, the law relating the tempo of finger movement to the phase shift does not explain why the phase shifts as tempo increases. The model is, in Robert Cummins's (2000) language, the description of an effect rather than its explanation. The model describes rather than explains the phenomenon.

Finally, the dynamicist might insist that the explanatory power of dynamicist models arises from their ability to describe general features about the behavior of systems independently of the material facts about the system that they describe. For example, the HKB model has been used to describe similar patterns of coordination between the two hands in a different bimanual pronation task (Carson et al. 2000), between the two arms in interlimb coordination tasks (Buchanan, Kelso, and DeGuzman 1997), between the arm and the leg (Kelso and Jeka 1992), between two individuals involved in coordinated social interaction (Schmidt, Carello, and Turvey 1990), and even between the gait transitions in horse locomotion (Schöner, Jiang, and Kelso 1990). One gets the sense that this is not coincidence and that the model is therefore picking up on a kind of pattern one witnesses in certain kinds of organized, coordinated systems generally. We agree that dynamical modeling is useful in part for revealing such widespread patterns. But we disagree that the scope of application

15. Not all counterfactuals are explanatory. Recall Lewis's (1986) important distinction between backtracking and nonbacktracking counterfactuals.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

for the model (that it applies to finger wagging, locomotion, etc.) affects its quality as an explanation. If we want to know why humans exhibit the phenomenon described in the HKB model, it is merely suggestive to note that a similar pattern is observed in a variety of other systems. This information might be useful in our search for general patterns in the organization of mechanisms, but it does nothing to explain the phenomenon we wanted to explain in the first place. If anything, it merely points out that many other similar phenomena require explanations as well, and perhaps these explanations will be similar. Whether they will in fact be similar is, of course, an open empirical question.

Many proponents of dynamic systems theory such as Kelso now appear to recognize the importance of mechanistic explanation. After developing the HKB model, Kelso and colleagues began researching how this behavioral regularity results from features of the underlying organization of component neural systems and their dynamics (see, e.g., Schöner and Kelso 1988; Jirsa et al. 1998; Jantzen et al. 2009). Kelso and colleagues (Jirsa et al. 1998) recently proposed a *neural field* model connecting the observed phase shift described by HKB to the underlying dynamics of neural populations in motor cortex. By mapping connections between the model components and components of neural systems, Kelso et al. have begun to transform a merely descriptive model into a mechanistic one. They seem to recognize the explanatory value of pushing beneath the regularities couched at the behavioral level to reveal explanatory mechanisms.

Dynamical models do not provide a separate kind of explanation subject to distinct norms. When they explain phenomena, it is because they describe mechanisms. As descriptive tools, they can be used to describe mechanisms phenomenally or mechanistically, correctly or incorrectly, and completely or incompletely.

5. The Difference-of-Gaussians Model of Visual Spatial Receptive-Field Organization. Above, we focused primarily on mathematical models of dynamical systems. Here, we consider mathematical models more generally, applying 3M to one of the most useful and well-known mathematical models in visual neuroscience: the *difference-of-Gaussians* (DOG) model. First introduced by Rodieck nearly 50 years ago, the DOG model (eq. [2]) describes the spatial receptive-field organization of retinal ganglion cells as a difference of two Gaussian functions.[16] The model has

16. A cell's *receptive field* is the circumscribed region of the retina within which illumination changes the cell's level of activation. *Receptive-field spatial organization* refers to how a cell's responsiveness or sensitivity changes across the spatial extent of its receptive field after visual stimulation (sometimes this is called the *receptive-field sensitivity function* or *spatial weighting function*).

https://doi.org/10.1086/661755 Published online by Cambridge University Press

become one of the most widely used and well-known mathematical models in visual neuroscience (e.g., Enroth-Cugell and Robson 1984; Shapley and Lennie 1985; Dayan and Abbott 2001). The model has since been extended to account for spatial receptive-field organization of visual cells in the dorsal lateral geniculate nucleus (dLGN) as well (So and Shapley 1981; Dawis et al. 1984).

Ganglion cells play an important role in early vision. They pool signals from multiple photoreceptors. In the process, they transform the graded membrane potentials that photoreceptors generate in response to light into an output signal in the form of trains of action potentials. These signals propagate to the dLGN, among other areas, in the form of spatiotemporal patterns of action potentials. The activity of individual retinal ganglion cells is determined largely by a population of photoreceptors sensitive to light coming from a circumscribed region of the visual field, known as the cell's receptive field. Kuffler (1953) delivered the first qualitative account of the receptive fields of ganglion cells. Quantitative characterizations followed soon after (Rodieck and Stone 1965a, 1965b; Enroth-Cugell and Robson 1966).

Kuffler pioneered the mapping of receptive fields by shining a small spot of light on a specific region of the retina while recording action potentials from individual ganglion cells. He showed that ganglion cells have circularly symmetric receptive fields arranged into two distinct concentric and mutually antagonistic regions (think of the bull's-eye and the next outermost region of a dartboard). For so-called ON-center cells, light on the center (the bull's-eye) causes the ganglion cell to fire (increase the rate at which it produces action potentials). Simultaneously shining light on the broader surrounding region of the receptive field diminishes the cell's response to light on the center. The OFF-center cells exhibit the opposite pattern, firing less when their center is illuminated and more when the surround is illuminated. Kuffler also found that the distribution of sensitivity (defined in terms of change in firing rate after stimulation) across both the receptive field center and surrounding regions are dome-shaped, antagonistic (or oppositely signed) sensitivities peaking at their shared middle and declining gradually toward their respective edges.

Rodieck (1965) was among the first to notice that these domelike distributions of sensitivity closely approximate Gaussian functions, particularly Gaussian responses graded over the distance from the absolute center of the ON-center cell's receptive field. Additionally, since the overall sensitivity profile of a ganglion cell across its entire receptive field involves center-surround antagonism, the difference operator is an obvious choice to express the mathematical relation between activation in these two regions. This is why Rodieck (1965) introduced the difference of two partially overlapping, circularly symmetric, and concentric Gaussian func-

https://doi.org/10.1086/661755 Published online by Cambridge University Press

tions as a mathematical model of the spatial receptive-field organization of retinal ganglion cells. According to the DOG model, the spatial receptive field of a given ganglion cell is expressed as the following function:

$$F(x, y) = \frac{A_1}{2\pi\sigma_1^2}\exp\left(-\frac{x^2 + y^2}{2\sigma_1^2}\right) - \frac{A_2}{2\pi\sigma_2^2}\exp\left(-\frac{x^2 + y^2}{2\sigma_2^2}\right), \qquad (2)$$

where the center of the receptive field is positioned at $(x, y) = (0, 0)$; the first term corresponds to the center component of the response (narrower, larger magnitude Gaussian), and the second subtracted term corresponds to the antagonistic or opposite-signed surround component (wider, smaller magnitude Gaussian); coefficient A_1 corresponds to the strength of the center response or, equivalently, to the peak local sensitivity of the center, at $(x, y) = (0, 0)$; A_2 corresponds to peak local sensitivity of the surround, also centered at (0, 0); and σ_1 and σ_2 are corresponding width parameters of the Gaussian envelopes of center and surround, respectively.

This model is a mathematically convenient and compact representation of the experimental data concerning receptive fields of neurons in early vision. The model has been confirmed with a variety of different stimuli, including larger circles, moving bars, and sinusoidal gratings (see, e.g., Enroth-Cugell and Robson 1966; Einevoll and Plesser 2005). And yet, in spite of its widespread use, notable predictive successes, and ability to represent accurately the target phenomenon, the DOG model does not explain why ganglion cells respond as they do to retinal inputs. The model was not even intended to do this.

Our reasons for thinking of the DOG model as a descriptive model parallel the reasons given above for both the Hodgkin-Huxley model and the HKB model.[17] First, the model is nothing more than a mathematically couched empirical description of the receptive field profiles of ganglion (and dLGN) neurons obtained from electrophysiological studies. The equation was generated to fit a particular set of experimental response curves and was not directly constructed to represent anything about the underlying retinal or geniculate circuit that could suitably explain why the receptive field has the particular spatial organization that it does. As with the Hodgkin-Huxley and HKB models that embody no commitments to the mechanisms that change membrane conductance or underlie bimanual coordination, the DOG model embodies no specific commitments concerning underlying neuronal morphology, physiology, and circuitry responsible for the ganglion cell receptive-field properties.[18] Rodieck does

17. For further discussion of this line of reasoning as it applies to the Hodgkin-Huxley model, see Craver (2006, 2007).

18. Except in the sense that whatever the underlying mechanism is, it must be capable of generating the behavior depicted by the DOG model, at least approximately.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

not constrain his mathematical description so that the variables in the model correspond to identifiable components, organizational features, and operations of the mechanism (as 3M demands). None of the variables or parameters in the DOG model correspond to components of the mechanism underlying the spatial receptive-field organization of early visual neurons.

How might one transform the DOG model from a descriptive, phenomenal model to an explanatory mechanistic model? The receptive field properties of ganglion cells and all cells downstream from them are determined by the anatomical connections between those cells and their upstream inputs. For example, an ON-center ganglion cell is activated by a circular spot stimulus because retinal photoreceptors centered over the ganglion cell's dendritic field are correspondingly illuminated. Ganglion cells have the receptive field properties they do because of the spatial arrangement of the photoreceptors that synapse directly onto the dendrites of the ganglion cell. The explanatory step will clearly involve understanding neuronal morphology and the synaptic connections in the retina and perhaps the developmental processes by which such functional relations are constructed, elaborated, and maintained. The goal is to show how the retinal components are organized together such that the behavior of the mechanism as a whole approximates the DOG model.

Visual neuroscientists continue to elaborate on the DOG model with the goal of bringing it more into line with 3M. Einevoll and Heggelund (2000), for example, construct a supplemented DOG model for dLGN receptive fields that incorporates what is known about the functional neuronal coupling in the geniculate circuit. Their model includes mathematical expressions for excitatory inputs received from a single afferent retinal ganglion cell (dLGN cells are known to receive direct excitation from one to several ganglion cells), and feed-forward inhibitory input signals received from intrageniculate interneurons, which in turn receive excitation from a few ganglion cells (Mastronarde 1992; Einevoll and Heggelund 2000). It depicts features of the underlying mechanism in a way that can be held accountable to the fruits of future inquiry into the underlying anatomy and physiology.

This example from visual cognitive neuroscience also illustrates the distinction between how-possibly and how-actually models. Cohen and Sterling (1991) consider three how-possibly models of the anatomical and electrophysiological circuits that could give rise to the Gaussian response curves described in the DOG model. The first model involves a given ganglion cell receiving afferent axonal projections from a single bipolar cell (a third variety of cell found in the retina) having a relatively wide dendritic field. This bipolar cell in turn receives axonal projections from a certain population of cone photoreceptors (around 100 or so neurons).

https://doi.org/10.1086/661755 Published online by Cambridge University Press

The Gaussian sensitivity profile is implemented in virtue of more cones from the middle of the field connecting to the bipolar cell than from those at the edge. The second how-possibly model involves afferent connections between a single ganglion cell and several bipolar cells, each of which has medium-width dendritic fields. These fields again receive direct input from the same population of retinal receptors. The Gaussian sensitivity in this model results from the fact that there are comparably more synaptic connections from the central bipolar cell onto the ganglion cell than from the other bipolar cells. Finally, the third model involves several bipolar cells, each with narrow dendritic fields. These fields in turn pool inputs only from the centermost group of cone photoreceptors in the overall population. The remaining cone cells supply signals indirectly via electrical interconnections to those centermost cones. In this model, the Gaussian sensitivity derives from this pattern of connectivity among photoreceptors.

Cohen and Sterling argue that only the third model is consistent with the available data about retinal microcircuitry. They rule out the first because the actual retinal circuitry involves an array of bipolar cells (rather than one). They rule out the second because the bipolar cells have narrow dendritic fields, and the bipolar array receives direct inputs from only a subset of the photoreceptor population (approximately 30 of the 90 cones in the population). The third model is validated because the anatomic circuit consists of an array of narrow-field bipolar cells that receive projections from about 30 cone photoreceptors with electrical coupling to the rest of the photoreceptor population. If they are right, the first two models are merely how-possibly mechanisms that have been excluded as candidate how-actually explanations. The third model purports to describe real components and organizational features of the mechanism that produces the retinal responses described by DOG. It is designed to satisfy 3M. It is a step in the direction of a more complete explanation.

Again, 3M is not merely an expression of imperialistic tendencies. The demand follows from the many limitations of predictivism and the conspicuous absence of an alternative model of explanation that satisfies scientific-commonsense judgments about the adequacy of explanations and does not ultimately collapse into the mechanistic alternative. We have relied, for example, on the idea that providing a general description of a phenomenon does not explain that phenomenon. Asked why the neuronal membrane changes its conductance to sodium and potassium ions as it does, it does not help to respond that all (properly functioning) neuronal membranes do so as well. For what we wanted to know was why neurons produce action potentials, and this answer tells us only that many neurons engage in some currently unexplained activity. Furthermore, the phenomenal description of the conductance change fails to reveal additional internal variables that might be used for the purposes of intervening to

https://doi.org/10.1086/661755 Published online by Cambridge University Press

change the behavior of the mechanism or to make better predictions about how the mechanism will behave in a variety of nonstandard conditions. We saw this in the case of the Hodgkin-Huxley model, and the same conclusions apply to the dynamical and cognitive models we have considered. Asked why a given ganglion cell changes its response to a moving spatial contrast grating, it is unhelpful to respond that all (properly functioning) ganglion cells behave in this way. This demonstration only multiplies the confusion. Moreover, the phenomenal description of the spatial receptive field falls short of disclosing underlying variables that might be used for the purposes of intervening to alter the response of the mechanism. The phenomenon characterized in the DOG model is an explanandum for visual cognitive neuroscience, not the explanans. To explain a phenomenon is not to show that it is consistent with the known regularities, but rather to show how it is situated in the causal structure of the world.

6. Conclusion. We defend a mechanistic approach to thinking about explanation at all levels of explanation in neuroscience, including systems and cognitive neuroscience. We demonstrate that widely accepted norms for evaluating explanations in neuroscience are best accounted for by the idea that to explain a phenomenon is to reveal its underlying mechanism. We show that precisely the same norms—captured by the 3M constraint—discern explanatory from nonexplanatory models of mechanisms across all levels of neural function. Finally, we show why neuroscientists should not content themselves with dynamical or other mathematical models that fail to satisfy 3M.

Those who hold that cognitive science and neuroscience must depart from the mechanistic tradition owe, in defense of their view, not only the reasons why mechanistic explanation is inappropriate in those domains but also a positive view of the intended, nonmechanistic form of explanation and the standards by which good nonmechanistic explanations are distinguished from bad nonmechanistic explanations. Absent such a positive view, the claim that nonmechanistic models explain is empty precisely because there are no clear standards for evaluating them. The onus is thus on advocates of nonmechanistic explanation in cognitive and systems neuroscience to make their view of explanation, and of the standards governing such explanations, explicit. Until such time, it would be wise to proceed on the assumption that the explanatory force of dynamical models, to the extent that they have such force, inheres in their ability to reveal dynamic and organizational features of the behavior of a mechanism.

REFERENCES

Baker, Jason M. 2005. "Adaptive Speciation: The Role of Natural Selection in Mechanisms of Geographic and Non-geographic Speciation." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:303–26.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

Bechtel, William. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.
Bechtel, William, and Adele Abrahamsen. 2002. *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks*. 2nd ed. Oxford: Blackwell.
———. 2005. "Explanation: A Mechanistic Alternative." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:421–41.
———. 2010. "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science." *Studies in History and Philosophy of Science* A 41:321–33.
Bechtel, William, and Robert C. Richardson. 1993/2010. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Repr. Cambridge, MA: MIT Press.
Bliss, Timothy V., and Terje Lomo. 1973. "Long-Lasting Potentiation of Synaptic Transmission in the Dentate Area of the Anaesthetized Rabbit following Stimulation of the Perforant Path." *Journal of Physiology* 232 (2): 331–56.
Bogen, James. 2005. "Regularities and Causality: Generalizations and Causal Explanations." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:397–420.
———. 2008. "The Hodgkin-Huxley Equations and the Concrete Model: Comments on Craver, Schaffner, and Weber." *Philosophy of Science* 75:1034–46.
Bressler, Steven, and J. A. Scott Kelso. 2001. "Cortical Coordination Dynamics and Cognition." *Trends in Cognitive Sciences* 5:26–36.
Buchanan, John J., J. A. Scott Kelso, and Gonzalo C. DeGuzman. 1997. "The Self-Organization of Trajectory Formation." Pt. 1, "Experimental Evidence." *Biological Cybernetics* 76:257–73.
Carson, Richard G., and J. A. Scott Kelso. 2004. "Governing Coordination: Behavioural Principles and Neural Correlates." *Experimental Brain Research* 154 (3): 267–74.
Carson, Richard G., Stephan Riek, Christopher J. Smethurst, Juan Franscisco Lisón Párraga, and Wintson D. Byblow. 2000. "Neuromuscular-Skeletal Constraints upon the Dynamics of Unimanual and Bimanual Coordination." *Experimental Brain Research* 131:196–214.
Chemero, Anthony, and Michael Silberstein. 2008. "After the Philosophy of Mind: Replacing Scholasticism with Science." *Philosophy of Science* 75:1–27.
Cohen, Ethan, and Peter Sterling. 1991. "Microcircuitry Related to the Receptive Field Center of the On-Beta Ganglion Cell." *Journal of Neurophysiology* 65 (2): 352–59.
Craver, Carl F. 2006. "When Mechanistic Models Explain." *Synthese* 153:355–76.
———. 2007. *Explaining the Brain*. Oxford: Oxford University Press.
———. 2008. "Physical Law and Mechanistic Explanation in the Hodgkin and Huxley Model of the Action Potential." *Philosophy of Science* 75:1022–33.
Craver, Carl F., and Anna Alexandrova. 2008. "No Revolution Necessary: Neural Mechanisms for Economics." *Philosophy and Economics* 24 (3): 381–406.
Cummins, Robert. 2000. "'How Does It Work' versus 'What Are the Laws': Two Conceptions of Psychological Explanation." In *Explanation and Cognition*, ed. Frank Keil and Robert A. Wilson, 117–45. Cambridge, MA: MIT Press.
Darden, Lindley. 2006. *Reasoning in Biological Discoveries: Mechanisms, Interfield Relations, and Anomaly Resolution*. New York: Cambridge University Press.
Darden, Lindley, and Carl F. Craver. 2002. "Strategies in the Interfield Discovery of the Mechanism of Protein Synthesis." *Studies in History and Philosophy of Biological and Biomedical Sciences* 33:1–28.
Dawis, Steven, Robert Shapley, Ehud Kaplan, and Daniel Tranchina. 1984. "The Receptive Field Organization of X-Cells in the Cat: Spatiotemporal Coupling and Asymmetry." *Vision Research* 24 (6): 549–64.
Dayan, Peter, and Larry F. Abbott. 2001. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press.
Douglas, Heather E. 2009. "Reintroducing Prediction to Explanation." *Philosophy of Science* 76:444–63.
Doyle, Declan A., João Morais Cabral, Richard A. Pfuetzner, Anling Kuo, Jacqueline M. Gulbis, Steven L. Cohen, Brian T. Chait, and Roderick MacKinnon. 1998. "The Struc-

https://doi.org/10.1086/661755 Published online by Cambridge University Press

ture of the Potassium Channel: Molecular Basis of K^+ Conduction and Selectivity." *Science* 280:69–77.
Dray, William. 1957. *Law and Explanation in History*. Oxford: Oxford University Press.
Einevoll, Gaute T., and Paul Heggelund. 2000. "Mathematical Models for the Spatial Receptive-Field Organization of Nonlagged X-Cells in Dorsal Lateral Geniculate Nucleus of Cat." *Visual Neuroscience* 17 (6): 871–85.
Einevoll, Gaute T., and Hans Ekkehard Plesser. 2005. "Responses of the Difference-of-Gaussians Model to Circular Drifting-Grating Patches." *Visual Neuroscience* 22 (4): 437–46.
Enroth-Cugell, Christina, and John G. Robson. 1966. "The Contrast Sensitivity of Retinal Ganglion Cells of the Cat." *Journal of Physiology* 187 (3): 517–52.
———. 1984. "Functional Characteristics and Diversity of Cat Retinal Ganglion Cells: Basic Characteristics and Quantitative Description." *Investigative Ophthalmology and Visual Science* 25 (3): 250–67.
Fuchs, Armin, Viktor K. Jirsa, and J. A. Scott Kelso. 2000a. "Issues in the Coordination of Human Brain Activity and Motor Behavior." *Neuroimage* 11 (5): 375–77.
———. 2000b. "Theory of the Relation between Human Brain Activity (MEG) and Hand Movements." *Neuroimage* 11 (5): 359–69.
Glennan, Stuart. 1996. "Mechanisms and the Nature of Causation." *Erkenntnis* 44:49–71.
Haken, Hermann, J. A. Scott Kelso, and H. Bunz. 1985. "A Theoretical Model of Phase Transitions in Human Hand Movements." *Biological Cybernetics* 51 (5): 347–56.
Haugeland, John. 1998a. "Mind Embodied and Embedded." In *Having Thought: Essays in the Metaphysics of Mind*, ed. John Haugeland, 207–40. Cambridge, MA: Harvard University Press.
———. 1998b. "The Nature and Plausibility of Cognitivism." In *Having Thought: Essays in the Metaphysics of Mind*, ed. John Haugeland, 9–46. Cambridge, MA: Harvard University Press.
Hedström, Peter, and Petri Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." *Annual Review of Sociology* 36:49–67.
Hempel, Carl G. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.
Hille, Bertil. 1992. *Ion Channels of Excitable Membranes*. 2nd ed. Sunderland, MA: Sinauer.
Hodgkin, Alan L. 1992. *Chance and Design: Reminiscences of Science in Peace and War*. Cambridge: Cambridge University Press.
Hodgkin, Alan L., and Andrew F. Huxley. 1952. "A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve." *Journal of Physiology* 117:500–544.
Jantzen, Kelly J., Fred L. Steinberg, and J. A. Scott Kelso. 2009. "Coordination Dynamics of Large-Scale Neural Circuitry Underlying Rhythmic Sensorimotor Behavior." *Journal of Cognitive Neuroscience* 21 (12): 2420–33.
Jirsa, Viktor K., Armin Fuchs, and J. A. Scott Kelso. 1998. "Connecting Cortical and Behavioral Dynamics: Bimanual Coordination." *Neural Computation* 10:2019–45.
Kelso, J. A. Scott. 1995. *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.
———. 2010. "Instabilities and Phase Transitions in Human Brain and Behavior." *Frontiers in Human Neuroscience* 4:23.
Kelso, J. A. Scott, Armin Fuchs, R. Lancaster, T. Holroyd, Douglas Cheyne, and Harold Weinberg. 1998. "Dynamic Cortical Activity in the Human Brain Reveals Motor Equivalence." *Nature* 392 (6678): 814–18.
Kelso, J. A. Scott, and John J. Jeka. 1992. "Symmetry Breaking Dynamics of Human Multilimb Coordination." *Journal of Experimental Psychology: Human Perception and Performance* 18 (3): 645–68.
Kitcher, Philip. 1989. "Explanatory Unification and the Causal Structure of the World." In *Minnesota Studies in the Philosophy of Science*, vol. 13, *Scientific Explanation*, ed. Philip Kitcher and Wesley Salmon, 410–505. Minneapolis: University of Minnesota Press.
Kuffler, Steven. 1953. "Discharge Patterns and Functional Organization of Mammalian Retina." *Journal of Neurophysiology* 16 (1): 37–68.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

Lewis, David. 1986. "Events." In *Philosophical Papers*, vol. 2, ed. David Lewis, 241–69. Oxford: Oxford University Press.
Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. "Thinking about Mechanisms." *Philosophy of Science* 67:1–25.
Mastronarde, David N. 1992. "Nonlagged Relay Cells and Interneurons in the Cat Lateral Geniculate Nucleus: Receptive-Field Properties and Retinal Inputs." *Visual Neuroscience* 8 (5): 407–41.
McDowell, John. 1996. *Mind and World*. Cambridge, MA: Harvard University Press.
Oullier, Olivier, Gonzalo C. DeGuzman, Kelly J. Jantzen, Julien Lagarde, and J. A. Scott Kelso. 2008. "Social Coordination Dynamics: Measuring Human Bonding." *Social Neuroscience* 3 (2): 178–92.
Piccinini, G., and Carl F. Craver. Forthcoming. "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches." *Synthese*.
Port, Robert F., and Timothy van Gelder. 1995. *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.
Rodieck, Robert W. 1965. "Quantitative Analysis of Cat Retinal Ganglion Cell Response to Visual Stimuli." *Vision Research* 5 (11): 583–601.
Rodieck, Robert W., and Jonathan Stone. 1965a. "Analysis of Receptive Fields of Cat Retinal Ganglion Cells." *Journal of Neurophysiology* 28 (5): 832–49.
———. 1965b. "Response of Cat Retinal Ganglion Cells to Moving Visual Patterns." *Journal of Neurophysiology* 28 (5): 819–32.
Rosenberg, Alex. 2001. "How Is Biological Explanation Possible?" *British Journal for the Philosophy of Science* 52:735–60.
Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.
———. 1989. "Four Decades of Scientific Explanation." In *Minnesota Studies in the Philosophy of Science*, vol. 13, *Scientific Explanation*, ed. Philip Kitcher and Wesley Salmon. Minneapolis: University of Minnesota Press.
Schmidt, Richard C., Claudia Carello, and Michael T. Turvey. 1990. "Phase Transitions and Critical Fluctuations in the Visual Coordination of Rhythmic Movements between People." *Journal of Experimental Psychology: Human Perception and Performance* 16: 247–77.
Schöner, Gregor, W. Y. Jiang, and J. A. Scott Kelso. 1990. "A Synergetic Theory of Quadrupedal Gaits and Gait Transitions." *Journal of Theoretical Biology* 142:359–91.
Schöner, Gregor, and J. A. Scott Kelso. 1988. "Dynamic Pattern Generation in Behavioral and Neural Systems." *Science* 239 (4847): 1513–20.
Shapley, Robert, and Peter Lennie. 1985. "Spatial Frequency Analysis in the Visual System." *Annual Review of Neuroscience* 8:547–83.
Skipper, Robert A., and Roberta L. Millstein. 2005. "Thinking about Evolutionary Mechanisms: Natural Selection." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:327–47.
So, Yuen T., and Robert Shapley. 1981. "Spatial Tuning of Cells in and around Lateral Geniculate Nucleus of the Cat: X and Y Relay Cells and Perigeniculate Interneurons." *Journal of Neurophysiology* 45 (1): 41–48.
Strevens, Michael. 2004. "The Causal and Unification Accounts of Explanation Unified—Causally." *Noûs* 38:154–76.
Thelen, Esther, and Linda Smith. 1994. *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.
Tognoli, Emmanuelle, and J. A. Scott Kelso. 2009. "Brain Coordination Dynamics: True and False Faces of Phase Synchrony and Metastability." *Progress in Neurobiology* 87 (1): 31–40.
van Gelder, Timothy. 1995. "What Might Cognition Be, If Not Computation?" *Journal of Philosophy* 92:345–81.
———. 1998. "The Dynamical Hypothesis in Cognitive Science." *Behavioral and Brain Sciences* 21:1–14.
van Gelder, Timothy, and Robert F. Port. 1995. "It's about Time: An Overview of the Dynamical Approach to Cognition." In Port and van Gelder, 1995, 1–43.

https://doi.org/10.1086/661755 Published online by Cambridge University Press

Walmsley, Joel. 2008. "Explanation in Dynamical Cognitive Science." *Minds and Machines* 18 (3): 331–48.
Woodward, James. 2003. *Making Things Happen*. New York: Oxford University Press.

https://doi.org/10.1086/661755 Published online by Cambridge University Press