Connectionism and Dynamical Explanation

Tim van Gelder
Department of Philosophy
Indiana University
Bloomington IN 47405
tgelder@ucs.indiana.edu

Abstract

A distinctive feature of connectionism as a research paradigm in psychology is use of a form of scientific explanation here termed dynamical explanation. In dynamical explanation, the behavior of a system is explained by reference to points and trajectories in an abstract state space. This paper contrasts dynamical explanation with some other major forms of scientific explanation, and discusses how dynamical explanation of the behavior of artificial neural networks can constitute genuine psychological explanation.

What is distinctive about connectionism as a research paradigm in psychology? This question has been approached from many different directions. Many have pointed to novel connectionist methods of representation; others, meanwhile, have focused on such issues as the connectionist emphasis on learning, the level at which it operates, whether or in what sense it employs rules, and so forth. Here I will be suggesting that, whatever else may be distinctive about it, connectionists are developing or at least, importing into psychology - a novel form of scientific explanation. Since this kind of explanation proceeds by constructing models conceived of as dynamical systems, the most appropriate term for this approach is dynamical explanation. In what follows I will give an intuitive introduction to the concept of dynamical explanation, contrasting it with other generic styles of explanation widely used in science, and will then discuss how dynamical explanation is deployed in connectionist modeling of psychological phenomena.

There is, of course, nothing new in the idea that connectionist networks can be thought of as dynamical systems, or in the idea that doing so is a key element in connectionism's distinctive perspective on cognition. (On both these points, see, e.g., (Smolensky, 1988).) I am not arguing for these points, but rather attempting to make explicit one form of scientific explanation one will be using if one describes networks as dynamical systems and uses such descriptions in accounting for psychological data. Making explicit the explanatory strategies in use is but one step in a general project of clarifying the conceptual foundations of connectionism.

Styles of Explanation.

In a well-known article outlining "cognitivism" (now more widely known as "classical" cognitive science), John Haugeland (Haugeland, 1981) pointed out that scientists commonly utilize at least three broad kinds of explanation - the deductive-nomological, the morphological, and the systematic. Deductive-nomological explanation is the classic; in this approach, scientists develop general, abstract, mathematically formulated laws, and then explain a given phenomenon by showing that it is merely an instance of the operation of such a law - i.e., by subsuming the phenomenon under the law. In morphological explanation, by contrast, an ability or disposition "is explained through appeal to a specified structure and to specified abilities of whatever is so structured" (p.247). For example, the uncanny ability of an eggshell to resist breaking when force is applied at the ends can be explained by pointing to the shape of the shell and the way such a shape distributes pressure. Here there is, at least in the first instance, no subsuming under laws; there are of course laws in the background, but knowledge of such laws is certainly not prerequisite for finding the explanation illuminating. In systematic explanation, the ability or disposition of some whole, in this case known as a "system," is also explained by pointing to a particular kind of internal structure, but this time one made up of distinct parts which participate in a "complexly organized pattern of interdependent interactions". For an example illustrating the systematic approach, consider an explanation of how a car engine works. Long before any laws become relevant, one has to describe the various parts of the engine (carburetor, cylinders, radiator etc.), how they individually function, and how each of their functions is integrated to produce a smoothly and powerfully rotating driveshaft.

These different explanatory strategies can be regarded as alternative ways of conceptualizing or "organizing" the world such that we can make sense of it. More than this, however: they tend to reflect real

differences in the way the world itself is constituted and events unfold. The difference between morphological and systematic explanation, for example, reflects the structural fact that systems, but not the kinds of mechanisms we describe morphologically, are made up of distinct parts which interact with each other in complex ways. In general, whether a particular explanatory strategy is the most appropriate in a given case depends on both the specific character of the situation under consideration and our explanatory interests and resources.

From this perspective, an obvious question arises: how many significantly different kinds of scientific explananda, and corresponding different explanatory strategies, are there? Or, less ambitiously: can we think of further kinds of situations, to which some further explanatory strategy is most appropriate? In the current context, of course, the ultimate goal of this line of questioning is cognition. What kind of mechanisms are cognitive mechanisms? Which explanatory strategy, or strategies, are the most appropriate in describing how cognition arises?

In fact, I think that there is at least one more major, generic form of explanation, one that is suggested by study of how explanations often proceed in distributed connectionist work, though it is most certainly not limited to that context. The kind of mechanism to which this species of explanation applies is the dynamical system. A dynamical system is any closed system whose state at a given time can be adequately captured by specifying the values of each of a set of parameters. As the dynamical system changes over time, the values of these parameters evolve in interdependent ways. In studying and explaining the behavior of dynamical systems one aims at formulating equations which describe the evolution of the system, and which can consequently be used to explain why the system is in the state it is in, or to predict what states it will come to be in. A classic example of a dynamical system is a pendulum. Parameters pick out the displacement and velocity of the bob, and relatively simple general equations govern how the values of these parameters change over time, capturing the periodic swinging motion of the pendulum.

The state space of a dynamical system is all the possible states which the system can be in. A state space can be represented using a vector space containing a point for every possible combination of values of all the relevant parameters; thus, the state space has as many dimensions as there are parameters. If we know the current state of the system i.e., the point in state space it currently occupies - then we can use the equations governing the behavior of the system to determine what point it will occupy next. A succession of such points is a trajectory in state space, and amounts to a picture of how the dynamical system changes as time goes on. Every point that the system might occupy lies on some trajectory or other, and shapes of the all the trajectories are fixed by the equations. Crucially, to understand how the system works is to have a sense for how the system changes over time, or, equivalently, to understand the general dynamical "topography" of the system. What do the trajectories look like? If the system is in a given state, how will it evolve?

In the case of the pendulum, the equations which tell us how the state of the system changes over time are given by general laws of classical mechanics. For this reason it might seem that dynamical explanations are just special instances of deductive-nomological explanation. In general, however, this is not true. For one thing, we can use dynamical explanations in situations where the equations governing the evolution of the system are not themselves general physical laws. In principle, of course, one supposes that the equations are ultimately derivable from general physical laws, and it may even be possible to provide such a derivation. However, the crucial point is that as long as one has the equations for the system, such a derivation is not in practice part of the actual explanation of the behavior of the system, and none of the explanatory force is lost if no such derivation is forthcoming.

Second, dynamical explanations may proceed without making explicit use of the equations governing the system. If the system is complex enough, one may not actually have the full equations in hand; alternatively, one might have them, but nevertheless find more perspicuous ways of explaining how the system works that proceed independently of the equations. This seems to be the situation on occasions when connectionists explain the performance of their networks by sketching a picture of the trajectories through which the overall patterns of activity evolve. By presenting a series of carefully selected two-dimensional snapshots of trajectories, they provide a sense of the dynamical structure of the system - i.e., an understanding of how the system behaves. Subsequent explanations of particular features of the behavior of the network can proceed by referring to that topography without needing to advert to the full equations which, formally, govern the behavior.

Dynamical explanations are not morphological or systematic either. In its pure form, dynamical explanation makes no reference to the actual structure of the mechanism whose behavior it is explaining. It tells us how the values of the parameters of the system evolve over time, not what it is about the way the system itself is constituted that causes those parameters to evolve in the specified fashion. It is concerned to explore the topographical structure of the dynamics of the system, but this is a wholly different structure than that of the system itself. This point is crucial to understanding dynamical explanations and how they differ from systematic explanations in particular. Dynamical explanations turn on the way the values of the parameters of the system interdependently evolve over time, whereas systematic explanations turn on the complex

interactions among the parts of the system itself. Consider the car engine again. Supposing a dynamical systems-style explanation were possible, it would begin by picking out the crucial parameters: engine temperature, r.p.m., gas level, timing advance, cylinder pressure - indeed, many of the quantities measured on by instruments on the dash, and no doubt a host of others besides - and then proceed to show how variations in one parameter affects the others, or how the engine typically proceeds through various states corresponding to regions of the state space (e.g., from cold to warm to out of gas), possibly according to rough equations describing the behavior of the particular engine. Explanations formed along these lines are quite different from systematic explanations which advert to the various parts of the engine and how they interact. In particular, the parameters utilized in dynamical explanations do not in general pick out parts of the system under study. Temperature and r.p.m. are not parts of the engine which interact with other parts. You can remove the carburetor leaving the rest of the engine behind, but you cannot remove the r.p.m..

The general point here is that dynamical explanation, which proceeds in terms of parameters, equations and state spaces, takes place at one level of remove from the actual mechanisms which produce the behavior quantified and explained in the dynamical account. To be sure, if one wanted an explanation of why one equation rather than some other governs a system, of why the state of the system travels through some trajectories and not others, one may be able to go on to offer a further, presumably systematic, explanation; this does not however make the initial dynamical explanations themselves either morphological or systematic, and the usefulness or validity of the dynamical explanation does not depend on one's being able to provide such further explanation.

I claim, then, that dynamical explanation constitutes a genuine alternative to other common forms of scientific explanation. The dynamical approach to explanation seems to have been neglected in the philosophy of science, but not in scientific practice itself, for a moment's reflection suggests that explanations fitting this general mold are widely used. What remains to be shown here is how dynamical explanation figures in connectionist work. This involves two steps: first, illustrating how the behavior of connectionist networks can be explained in dynamical terms; and second, showing how dynamical explanation of connectionist networks combines with the technique of modeling to construct genuinely psychological explanations.

Dynamical Explanation of Network Behavior.

It is probably obvious enough how connectionist networks are conceived as dynamical systems. In the simplest and most familiar case, the parameters specify activity levels for each of the units, and equations based on the processing characteristics of the units and the way they are interconnected describe how activity levels will evolve over time. Patterns of activation over the network as a whole can be regarded as points in an activation state space, and processing can be thought of as traveling along a trajectory in activation space. From this perspective, to explain the behavior of a network is to provide a sense for the dynamical structure of the network as a whole, such that any particular state it might be in can be related to other states that it was or will be in; or such that one trajectory can be contrasted with others. This "sense" can be obtained either formally, by providing equations capturing the dynamical structure, or intuitively, by (for example) sketching trajectories in a selected few of the dimensions. (Note that explanation of the behavior of the network in terms of its dynamical structure must be carefully distinguished from another common explanatory task in connectionism, namely, explaining how a network comes to have the dynamical structure that it does.)

Hopfield nets (Hopfield, 1982) and Boltzmann machines (Hinton and Sejnowski, 1986) are excellent examples of connectionist networks whose behavior is naturally understood in dynamical terms. In general it is helpful to see processing, over many time steps, as tracing out a somewhat erratic trajectory in the activation state space of the network. Indeed, if we add to the space a dimension for another parameter such as global energy, the settling process characteristic of such networks can be visualized as a downward slide to a resting place. Explanations of particular aspects of the networks behavior typically refer to the general structure of these dynamics. Why did the network settle at a particular point? Because it began its settling process at another point which happened to be in the basin of attraction for the settling point.

Many other connectionist explanations proceed in basically dynamical terms, though the dynamical character is often concealed by the fact that they are often highly simplified varieties of the strategy. Explanations in this latter category tend to consider only a few time steps (i.e., highly truncated "trajectories") and, from step to step, shift attention from one sub-space of the overall activation space to another. Consider for example the explanation of generalization phenomena in the well-known past tense learning model of Rumelhart & McClelland (Rumelhart and McClelland, 1986). The ability of the network to produce the correct past tense form of an unseen, irregular verb such as *bid* is explained in terms of the fact that the trained network is "sensitive to the subregularities as well". That is, there are regularities even among irregular verbs; irregular verbs with similar present tense forms often also have similar past tenses. To say that the network is "sensitive to the subregularities" is to say that it treats members of these subgroups in a similar fashion. Thus, explaining the network's correct performance on

the verb *bid* is a matter of (a) pointing out that its input representation is similar (i.e., close in the space of input patterns) to other irregular verbs, and (b) adverting (in this case, in a highly informal way) to the overall dynamical structure of the network. This explanation considers a "trajectory" of only one time step and its description of the topological structure of the dynamics of the network is framed only in terms of how points in the input-unit activation sub-space are mapped to the output-unit subspace.

A particularly provocative example of dynamical explanation at work is found in Jeff Elman's descriptions of his SRN models of sentence processing (see e.g., (Elman, 1989)). Elman made effective use of an increasingly common technique exploiting principal components analysis to select an appropriate two-dimensional "window" onto the activation state space - or rather, in this case, the sub-space corresponding to hidden unit activity. In this window one plots a series of points, corresponding to states that the hidden units go through upon presentation of a series of inputs; this series of points amounts to a picture of the activation trajectory for those inputs. Collectively, a series of such trajectory pictures yields at least a glimmer of understanding of the complex dynamical structure of the network. Explanations of the successful performance of the network at its word prediction task make crucial reference to these trajectories; thus, the network is held to be able to correctly predict the next word in a complex sequence precisely because the structure of the sequence up to that point had been encoded in the activation trajectory (see (Port and van Gelder, 1991)

Why are connectionists increasingly using dynamical explanation in preference to, for example, any of the other forms of explanation described by Haugeland? The quick and easy answer is that neural networks are themselves the kinds of systems for which the most natural explanations are dynamical. Deductive-nomological explanation requires general covering laws, yet the systems of differential equations governing the behavior of complex networks are not general laws, and if they can be formulated at all for such networks, are typically unwieldy and unilluminating. Morphological explanation is inappropriate since it does not make room for the change and interdependence of the parameters; and systematic explanation is of little help since the "parts" of the structure are so many and so similar, and key parameters (e.g., "energy") do not refer to parts of the system at all. In dynamical explanation one abstracts away from any consideration of how the system under study is actually put together, and focuses only on how various parameters change in interdependent ways. It seems that this distancing from the implementation is an essential simplifying step in attaining a deep understanding of how these particular kinds of highly complex systems work.

Dynamical Explanation in Psychology.

These examples of dynamical explanation in connectionism only involve explanation of the behavior of networks themselves; psychological data has not yet entered the story. How then can dynamical explanation constitute psychological explanation?

To answer this question, the concept of dynamical explanation of neural networks has to be combined with an understanding of how those networks are supposed to function as models of psychological phenomena. In modeling we explore one relatively familiar, often artificial structure or mechanism as a means of exploring another less familiar one. Thus, in psychology, and speaking in the broadest possible terms, one thing we wish to do is understand that mechanism which produces our behavior - at least, our behavior as codified in the data of experimental psychology. That mechanism, of course, is the brain. (Some prefer to distance themselves more from the messy neurobiological details, and so call the mechanism the mind.) Yet that mechanism is so awesomely complex, and the gap between neuroscience and psychology still so large, that we cannot yet explain how the brain (or mind) produces our behavior directly. As an intermediate step, we construct a model - a substitute mechanism which is supposed to be relevantly similar to the original, at least so far as the latter is described at some suitably abstract level. By exploring and understanding the properties of the model, we hope to reach some understanding of the original mechanism.

An obvious test of the adequacy of a model is that it account for the data - that is, it should produce the same overall behavior as the mechanism that is the ultimate explanatory target. Most psychologists producing models of human performance direct virtually all their attention at the constraints provided by this requirement. In principle, however, we can judge the adequacy of the model by reference to any knowledge we happen to have of the original mechanism. For example, a model of internal processes in the sun can be judged not only according to whether it generates the right behavior (corresponding to light and heat output, flares, sunspots etc.) but also according to whether it relies on processes that accord with other quite general knowledge from chemistry and physics. In the psychological case the "top-down" constraints of matching performance can, at least in principle, be supplemented with "bottom-up" constraints provided by our increasing knowledge of the brain, its structure and modes of functioning.

No model will be identical, in all its features, to the mechanism that is to be explained; only some aspects of the model are relevant to whether it meets the constraints provided by the performance data, and only some aspects are relevant to whether it accords with neuroscientific evidence of the basic constitution of the neural mechanism. Thus, whenever a model is proposed, it ought to be accompanied by what Wilfred

Sellars called a commentary, specifying which features of the model are to be taken as relevant to whether it satisfies the various constraints. For example, implicit in the discussion of the past tense learning model was the claim that which set of Wickelfeatures is output for a given input is taken to be relevant to whether the model accounts for the psychological data, while how fast it produces that set is not relevant. Other things being equal, a better model is one in which more of its features are judged relevant to its acceptability as a model.

Now, connectionist explanations of the behavior of their networks in dynamical terms constitutes dynamical explanation in psychology if (a) the network is being used as a model for the mechanism producing the psychological data, and (b) it is explicitly claimed or at least implicitly assumed that the abstract dynamical structure of the network corresponds to the abstract dynamical structure of the original mechanism. If these conditions are satisfied, then a connectionist's explanation of the performance of her network in terms of locations and trajectories in unit activation space is, at the same time, explanation in dynamical terms of how humans exhibit the performance they do. Note that for connectionist explanations to count as genuine dynamical explanations of psychological data, it is not required that the processing units of the model be interpreted as corresponding to real neurons, the connections as real synapses, etc.; the commentary may simply leave the interpretation of the units themselves undetermined. What matters is the overall structure of change, not what it is that is changing.

In short, connectionists are, increasingly, producing dynamical explanations of psychological data by combining dynamical explanation of the behavior of their networks with a certain way of interpreting those networks as models of the actual mechanisms underlying human performance. It may eventually be possible to produce dynamical explanations in psychology without first producing and explaining artificial neural network models; until the actual mechanisms are much better understood, however, modelling of some kind is an essential intermediate stage.

Conclusion.

According to Haugeland, understanding how mainstream symbolic cognitive science uses systematic explanation is essential to understanding its distinctive approach to cognition and how it can constitute respectable scientific investigation in psychology. It would probably be premature to assert, at this stage, that the concept of dynamical explanation will play a similar role for connectionism. It is safe to say that dynamical explanation is increasingly common in connectionist practice, arising in response to at least two pressures the increasing complexity of connectionist models, and the developing interest in the temporal structure of cognitive processes. I am not claiming that the use of dynamical explanation is any kind of criterion that can be used to distinguish connectionist approaches from others; in general, it is a methodological blunder to suppose that there is any fail-safe distinguishing marker. Further, I am not claiming that dynamical explanation is the only kind of explanation connectionists use. It is, nevertheless, a distinctive and novel feature of the connectionist perspective on cognition.

References

Elman, J. (1989). Representation and structure in connectionist models. Technical Report 8903, Center for Research in Language, UCSD, La Jolla, CA 92093-0108.

Haugeland, J. (1981). The nature and plausibility of cognitivism. In Haugeland, J., editor, *Mind Design*, pages 243–281. Bradford/MIT Press, Cambridge, Mass.

Hinton, G. and Sejnowski, T. (1986). Learning and unlearning in boltzmann machines. In Rumelhart, D. and McClelland, J., editors, *Parallel Distributed Processing, Vol. 1*, pages 282–317. The MIT Press, Cambridge, MA.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences*, volume 79, pages 2554–2558. National Academy of Sciences.

Port, R. and van Gelder, T. (1991). Representing aspects of language. In *Proceedings of the 13th Meeting of Cognitive Science Society*, Hillsdale, NJ. L. Erlbaum Assoc.

Rumelhart, D. and McClelland, J. (1986). On learning the past tense of English verbs. In McClelland, J. and Rumelhart, D., editors, *Parallel Distributed Processing*, volume 2, pages 216–271. MIT Press, Cambridge, MA.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74.