From connectionism to eliminativism

Stephen P. Stich

Department of Philosophy, University of California, San Diego, La Jolla, Calif. 92093

Smolensky's portrait of connectionism is a welcome and exciting one. The burden of my commentary will be that if the project he describes can be carried off, the consequences may be much more revolutionary than he suggests. For if it turns out that Smolensky-style connectionist models can indeed be constructed for a broad range of psychological phenomena of both the "intuitive" and the "consciously conceptualized" sort, then, it seems to me, a pair of very radical conclusions will plausibly follow. The first is that folk psychology – the cluster of commonsense psychological concepts and principles that we use in everyday life to predict and explain each other's behavior – is in serious trouble. The second is that much psychological theorizing that cleaves to what Smolensky calls the "symbolic paradigm" is in serious trouble as well. In both cases, the trouble I envision is the same: The theories are false and the things they posit don't exist. Since space is limited, I'll limit my remarks to theories in the symbolic paradigm, and leave folk psychology for another occasion.

A central thesis in Smolensky's rendition of connectionism is that a "complete, formal account of cognition" does not "lie at the conceptual level" but at the "*sub*conceptual level" (sect. 2.4, para. 5). Earlier, in making much the same point, he tells us that "complete, formal and *precise* descriptions of the intuitive processor are generally tractable not at the conceptual level, but only at the subconceptual level" ((8)c, sect. 2.3, para. 7, emphasis added). But what exactly does Smolensky have in mind when he claims that a complete, formal, *precise* account of cognition is to be found only at the subconceptual level? As I read him, what Smolensky is claiming is that the real, exceptionless, counterfactual supporting generalizations or laws of cognition are only to be found at this level. At the conceptual level, by contrast, such generalizations as we have will be at best rough and ready approximations that may be more or less accurate within a limited set of boundary conditions, and generally not very accurate at all when we go outside those boundary conditions. If this thesis turns out to be correct, then the cognitive states and processes posited by connectionist models will be the ones describable by genuine laws of nature, but there will be no laws describing the doings of the semantically interpreted mental symbols posited by theories at the symbolic level. If we want accurate predictions of the phenomena, they will have to be sought at the subsymbolic level. As Smolensky would be the first to agree, the thesis he sketches is at this point only a hopeful guess. To defend it requires that connectionists actually build models for a broad range of phenomena, and demonstrate that they do indeed yield more accurate predictions than competing models at the conceptual level. But let us assume that the thesis will ultimately be established, and consider the consequences for theories and posits at the conceptual level.

To start us off, an analogy may prove helpful. For Lavoisier, in the last quarter of the 18th century, heat was caused by caloric, an "exquisitely elastic fluid" "permeating all nature, which penetrates bodies to a greater or lesser degree in proportion to their temperature" (quoted in Gillispie 1960, p. 240 & p. 239). When Sadi Carnot formulated the second law of thermodynamics in 1822, he "still handled caloric as flowing from a real reservoir of heat down a continuous gradient" (Gillispie 1960, p. 241). For many years the theory of heat that posited caloric was embedded in an evolving, progressive, sophisticated research program that generated both explanations of observed phenomena and increasingly accurate predictions. Ultimately, however, that theory was rejected and replaced by the kinetic theory. Though the detailed history of this transition is a complicated story, a crucial factor was that the new theory sustained more accurate predictions and better explanations over a broader range of phenomena. Moreover, since the kinetic theory posits no "exquisitely elastic fluid," and recognizes no laws governing its flow, those who were prepared to grant that the kinetic theory is better concluded that caloric theory is false, and that the fluid it posits does not exist.

Consider now the analogies that will obtain between this case and the case of conceptual level psychological theories if Smolensky's thesis turns out to be right. Like caloric theory, the conceptual paradigm has sustained an evolving, progressive, sophisticated research tradition. But if Smolensky is right, we will find that the generalizations of conceptual level theories (like those of caloric theory) are only approximations and apply only in limited domains, while the generalizations of subconceptual level theories (like those of kinetic theory or its successors) are "complete" and "precise." Against the background of this analogy, it is tempting to conclude that if Smolensky's thesis is right, then conceptual level theories are false, and the entities they posit do not exist.

There is reason to suppose that Smolensky himself would not resist the first half of this conclusion. For at one point he tells us that "the relationship between subsymbolic and symbolic models is . . . like that between quantum and classical mechanics" (sect. 5, para. 11). But, of course, if quantum mechanics is right, then classical mechanics is wrong. Whatever its virtues, and they are many, classical mechanics is a false theory.

The second half of the conclusion I'm trying to coax from my analogy is the more distinctively eliminativist half. (For some background on "eliminativism" see P. M. Churchland 1894, pp. 43–49; P. S. Churchland 1986, pp. 395–99; Stich 1983, Chapter 11.) What it claims is that the entities posited by conceptual level theories are like caloric in one very crucial respect; they do not exist. From his one brief mention of "naive . . . eliminative reductionism" (sect. 10, para. 2). I'd guess that Smolensky would be more reluctant to endorse this half of my conclusion. Nor would such reluctance be patently unjustified. For it is certainly not the case that whenever one theory supplants another we must conclude that the entities posited by the old theory do not exist. Often a more appropriate conclusion is that the rejected theory was wrong, perhaps seriously wrong, about some of the properties of the entities in its domain, or about the laws governing those entities, and that the newer theory gives us a more accurate account of those very same entities. Thus, for example, pre-Copernican astronomy was very wrong about the nature of the planets and the laws governing their movement. But it would be something of a joke to suggest that Copernicus and Galileo showed that the planets Ptolemy spoke of do not exist. So to defend the eliminativist half of my conclusion, I must argue that the connectionist revolution, as Smolensky envisions it, bears a greater similarity to the rejection of the caloric theory than to the rejection of geocentrism.

In arguing the point, it would be useful if there were, in the philosophy of science literature, some generally accepted account of when theory change sustains an eliminativist conclusion and when it does not. Unfortunately, however, there is no such account. So the best we can do is to look at the posits of the old theory (the ones that are at risk of elimination) and ask whether there is anything in the new theory that they might be identified with. If the posits of the new theory strike us as deeply and fundamentally different from those of the old theory, in the way that molecular motion seems deeply and fundamentally different from "exquisitely elastic" caloric fluid, then the eliminativist conclusion will be in order. Though, since there is no easy measure of how "deeply and fundamentally different" a pair of posits are, our conclusion is bound to be a judgment call. That said, let me offer a few observations which, I think, support a proeliminativist judgment.

Smolensky notes, quite correctly in my view, that in the

dominant approach to cognitive modeling (the approach that he calls the "symbolic paradigm") symbols play a fundamental role. He goes on to note that these symbols have a pair of fundamental characteristics: They refer to external objects and they are "operated upon by 'symbol manipulation'" (cf. sect. 1.3., para. 3). Smolensky does not elaborate on the idea that symbols are operated on by symbol manipulation, but I take it that part of what he means is that, in the models in question, symbol tokens are assumed to have a reasonably discrete, autonomous existence; they are the sorts of things that can be added to, removed from or moved around in strings, lists, trees and other sorts of structures, and this sort of movement is governed by purely formal principles. Moreover, in the symbolic paradigm, these sorts of symbol manipulations are typically taken to be the processes subserving various cognitive phenomena. Thus, for example, when a subject who had previously believed that the hippie touched the debutante comes to think that the hippie did not touch the debutante, symbolic models will capture the fact by adding a negation operator to the discrete, specifiable symbol structure that had subserved the previous belief. Similarly, when a person acquires a new concept, say the concept of an echidna, symbolic models will capture the fact by adding to memory one or more symbol structures containing a new, discrete, independently manipulable symbol that refers to a certain class of external objects, namely echidnas.

In connectionist models, by contrast, there are no discrete, independently manipulable symbols that refer to external objects. Nor are there discrete, independently manipulable clusters of elements (or "subsymbols") which may be viewed as doing the work of symbols. When a network that had previously said yes in response to "Did the hippie touch the debutante?" is retrained to say no, it will generally not be the case that there is some stable, identifiable cluster of elements which represent the proposition that the hippie touched the debutante, both before and after the retraining. And when a network that was previously unable to give sensible answers to questions about echidnas is trained or reprogrammed to give such answers, there typically will not be any identifiable cluster of elements which have taken on the role of referring to echidnas. Instead, what happens in both of these cases is that there is a widespread readjustment of weights throughout the network. As Smolensky notes, the representation of information in connectionist models (particularly in parallel distributed processing style models) is widely distributed, with each unit participating in the representation of many different aspects of the total information represented in the system. This radical disparity between strategies of representation in symbolic and PDP models makes a smooth reduction – or indeed *any* reduction – of symbols to elements (or to patterns of activity) extremely implausible. Rather, I submit, the relation between mental symbols and connectionist elements (or patterns of activity) is akin to the relation between caloric and molecular motion. If this is right, then in those domains where connectionist models prove to be empirically superior to symbolic alternatives, the inference to draw is that mental symbols do not exist.

From data to dynamics: The use of multiple levels of analysis

Gregory O. Stone

Department of Psychology, Arizona State University, Tempe, Ariz. 85281

While focusing on the substantive differences between connectionism and traditional cognitive science, Smolensky's analysis illustrates a fundamental epistemological difference. In the traditional approach, the symbolic level is the "correct" level of analysis. Other levels, such as hardware implementation, are effectively considered irrelevant. In contrast, Smolensky argues that "successful lower-level theories generally serve not to replace higher-level ones, but to enrich them, to explain their successes and failures, to fill in where the higher-level theories are inadequate, and to unify disparate higher-level accounts." Thus, connectionism may portend a revival, in cognitive science, of theoretical pluralism – the philosophy that no single perspective can fully account for observed phenomena (James 1967).

Smolensky presents three levels of analysis (neural, subconceptual, and conceptual) as a priori theoretical constructs. I will argue, however, that the choice of levels derives from a strategy of maximizing the explanatory power of the pluralistic framework in which they are embedded. In other words, levels of analysis are primarily pragmatic constructs.

What are the advantages of a pluralistic methodology? One common objection to connectionist models is that their complexity hinders an understanding of what they are doing and why. This conceptual opacity is, to some extent, a price paid for their flexibility and generality of application, allowing models built from a few basic mechanisms to account for a broad range of disparate phenomena. On the other hand, mechanisms explicitly tailored for specific operating characteristics tend to be limited in their generality. This often leads to a profusion of unconnected, but eminently testable and transparently interpretable, special-purpose models. A methodology which uses and interrelates both levels of analysis can exploit the strengths and overcome the weaknesses of each when considered in isolation.

A concrete example will help to clarify this point. Reeves and Sperling (1986) asked subjects to report, in order, the first four items (digits) from a rapidly presented visual sequence. But subjects first had to shift their attention from another part of the visual field to the digit stream. The attention shift altered the perceived order of items in the sequence, producing an inverted-U shaped recall function. In the first phase of their analysis, they found that a scalar precedence or order score for each item in each condition provided a very powerful account of the data. However, this analysis invoked a large number of parameters and offered no conceptual insight into why the observed precedences were obtained. Their second level of analysis produced a close fit to these precedence scores by treating them as the result of the temporal integration of an item's input strength. A slow-opening attention gate reduced the input strength of early items, which lead to the inverted-U shape of precedence scores across position. This level of analysis provided a conceptual framework with greater parsimony; however, it was domain-specific and provided no link to temporal order in short-term memory in the absence of an attention shift.

Grossberg and Stone (1986) extended the analysis to the subconceptual level by mapping the Reeves and Sperling model into the short-term memory dynamics of adaptive resonance theory. The analysis began with an abstraction from the extremely complex activation dynamics to an emergent and more tractable functional form relating the relative precedence strengths. This emergent functional form is necessary for stable, long-term encoding. When this functional form was applied to the Reeves and Sperling model, several unexpected principles of short-term memory dynamics and attentional gain control were revealed. Furthermore, the experimentally derived order scores were accounted for using a mechanism which plays a critical role in adaptive resonance theory treatments of other short-term memory phenomena, as well as treatments of categorization, unitization, and contextual facilitation (see articles reprinted in Grossberg 1987a; 1987b). The key point in this example is that it would have been difficult – if not impossible – to have achieved the same degree of insight by mapping the data directly into the class of possible short-term memory dynamics.

Each level of analysis in the preceding example served an important role in the overall methodology.

The descriptive level of analysis encapsulates the raw data in a