3 What is the "D" in "PDP"? A Survey of the Concept of Distribution

Tim van Gelder
Indiana University

Suppose there were such a thing as the "computational theory of mind" (CTM); and suppose that, for whatever reason, you were dissatisfied with it. You may well be tempted to ask: What would an alternative look like? Could there be an alternative that was even remotely plausible? Is connectionism in the business of developing such an alternative?

With issues such as these in vogue recently, considerable attention has been given to the preparatory task of succinctly characterizing some version of CTM to which the desired alternative can stand opposed. One point of universal consensus has been that an essential feature of CTM is the use of symbolic representations. Any theory failing to employ such representations automatically falls outside the broad CTM umbrella. This suggests an obvious approach to the questions just raised. Assuming that any remotely plausible theory of mind must be based on manipulation of internal representations of some kind, we need to find some other generic form of representation to play a foundational role in the new theory analogous to that played by symbolic representation in CTM. Having found such a form, we could evaluate the general plausibility of a theory of mind constructed around it. Perhaps connectionist work contains some clues here, both about the form itself and about the kind of theory in which it would be embedded.

The alternative form of representation required by this approach has to satisfy some demanding conditions. It must, of course, be demonstrably nonsymbolic, but it must also be sufficiently general to allow the characterization of a reasonably broad conception of the mind. This means, among other things, that it must be rich enough to encompass a wide variety of particular articulations (just as symbolic representation can be instantiated in a very wide variety of particular ways), and yet characterizable in a way that is sufficiently abstract to transcend all kinds of irrelevant implementation details. Crucially, it will have to be powerful enough to make possible the effective representing of the kinds of information that are essential to human cognitive performance. Preferably, this alternative will have some deep connection with neural network architectures, thereby minimizing future difficulties relating the new theory to the neurobiological details, and in the meantime allowing us to both interpret and learn from connectionist research.

This is a tall order by any account, and a moment's reflection reveals that there are few if any plausible candidates available. The traditional cognitive science literature is of remarkably little help here. In that relatively small portion concerned specifically with the actual form of mental representation, symbolic styles have for the most part been contrasted only with broadly imagistic styles (pictorial, analog, etc.). For a long period the most notable research taking a manifestly nonsymbolic approach was the investigation of mental imagery, and surveys of the field typically treat these two broad categories as the only relevant possibilities. Yet, although the category of imagistic representations might begin to satisfy some of the constraints just listed, it clearly fails to satisfy others; in particular, it is generally accepted that imagistic representations are not powerful enough to underlie central aspects of cognition such as linguistic performance and problem solving. There has even been serious debate over whether mental imagery itself is strictly imagistic.

One response to this apparent lack of plausible alternatives is to accept that representations must be symbolic in some suitably generic sense, and consequently to maintain that any feasible alternative to CTM must differ not in how knowledge is represented but rather in how the representations themselves are manipulated—that is, in the nature of the mental processes. Yet this approach also is unpromising. Representations and processes tend to go hand in hand; the way knowledge is represented largely fixes appropriate processes and vice versa. For this reason, conceding that representations must be generically symbolic places one in a conceptual vortex with the standard CTM at the center.

One reason there appear to be so few alternatives is that the conception of symbolic representation invoked in characterizations of CTM is so very general, and usually rather vague. This suggests a more cautious gambit: fine-tune the conception of symbolic representation itself, articulating some more specific formulation that can fairly be attributed to CTM, thereby making room for some quasi-symbolic alternative between analog anarchy on one hand and the rigors of strictly syntactic structure on the other. However, although headway can certainly be made in this direction, it has an obvious strategic flaw: Major differences in paradigms are unlikely to rest on delicate philosophical distinctions, and if perchance they did, it would be relatively difficult to convince others of the fact. It is vastly preferable to propose a style of representation with unquestionable antisymbolic credentials. If there actually is any quasi-symbolic option of the kind just mentioned, it should be introduced as a special case of a manifestly distinct category, rather than as some subtle variant on the standard symbolic model.

If at this point we look to connectionism, it is difficult to avoid noticing the frequent emphasis on distributed representation, an emphasis evident even in the familiar designation "parallel distributed processing" (PDP). Distributed representation may well satisfy the first of the requirements on an acceptable alternative, because it is often deliberately contrasted with symbolic representation (e.g., as when it is claimed that a network knows how to form the past tense without the benefit of explicit symbolic rules). Moreover, the category appears to be appropriately general; at one time or another, distributed representation has cropped up in areas as diverse as functional neuroanatomy, psychology of memory, image processing, and optical phenomena such as holography; indeed, researchers originally began applying the term *distributed representation* in connectionist contexts precisely because of perceived similarities between connectionist representations and these other cases. Considerations such as these suggest we should inquire into the possibility that distributed representations form the kind of category we are after. Perhaps, in other words, there is here a natural kind of representation, a kind that includes all or most of the cases previously described as distributed, whose members are somehow inherently nonsymbolic, but that is nevertheless sufficiently rich, powerful, and so on, that it might form the basis of some plausible alternative to CTM.

The immediate difficulty with this suggestion is the lack of any clear account of what distributed representation actually is. The concept itself is relatively novel, and though many people have recently offered their preferred brief characterizations, it has had almost no serious treatments as an independent topic of investigation.[1] Worse, there is very little consensus even in such characterizations as are available. The diversity of definitions suggests that there really is no unified category of distributed representations after all. Feldman (1989) for one has concluded ". . . people have been using the term ["distributed"] to denote everything from a fully holographic model to one where two units help code a concept; thus, the term has lost its usefulness (p. 72). Clearly, before we can even begin to take seriously the idea that a plausible alternative to CTM might be constructed on a distributed foundation, we need to formulate a reasonably clear and comprehensive account of the nature of distributed representation. This task goes vastly beyond what might be achieved here; what follows is simply an exploratory overview of the current concept (or concepts), a disentangling of some of the many themes and issues that have at one time or another been associated with distribution.

[1] A notable exception here is work by Walters (see, e.g., Walters, 1987), although her concerns are considerably more restricted than those taken up here. A preliminary account of distributed representation in connectionist contexts is found in Hinton, McClelland, and Rumelhart (1986).

DIVERSE DEFINITIONS OF DISTRIBUTION

A useful point of entry is to note the inadequacy of one style of definition common in connectionist work. In perhaps the most authoritative version (Hinton, McClelland, & Rumelhart, 1986), representations are alleged to be distributed if: "Each entity is represented by a pattern of activity distributed over many computing elements, and each computing element is involved in representing many different entities" (p. 77).[2] The most obvious problem here, from the current perspective, is one of narrow focus. A distributed representation is defined as a "pattern of activity" over "computing elements" specifically; but this is too limited even for connectionist purposes, because there at least two species of distributed representation in connectionist networks—the patterns of activity themselves, and the patterns of connectivity that mediate their transformation. It may be that these are in fact essentially interlocked, each needing the other, but there is at least a prima facie distinction, because the two kinds of representation appear to have some significantly different characteristics. Thus this definition would have to be generalized significantly if it were to capture the notion of distribution implicit even in connectionist work, let alone whatever is common to cases as diverse as those mentioned above.

Narrowness is not however the worst of its problems. The intended contrast is with a variety of "localist" representation in which each entity is represented by activity in a single computing element. But in its concern to distinguish distribution from these kinds of localist cases, this definition patently fails to distinguish it from other cases that, surely, are not distributed in any interesting sense. For many familiar kinds of representation count as "patterns of activity" over sets of "computing elements" ("units," "locations," or whatever); in particular, when numbers are encoded as strings of bits in a register of an ordinary pocket calculator, they are being represented by distinctive activity patterns, and each unit or location participates in the representing of many different numbers over the course of a calculation.[3] This leaves two possibilities: either distribution is not an interestingly distinct category after all; or it is, but one whose essence eludes this definition. The latter turns out to be vastly more fertile as a working hypothesis.

Day to day practice often compensates for deficiencies in overt formulation. The real content of this characterization is implicit in the way it is received and guides construction of new connectionist schemes of representation. In this light, the central theme of this version—the representing of entities as different patterns of activity over groups of units—deserves closer scrutiny. Consider first the

[2] Here, to avoid obvious circularity, we should read "distributed" as meaning something like "spread over."

[3] Single bit storage locations in a digital memory can be seen as extremely simple computing units, changing state according to their inputs, and outputting their state when accessed. See McEliece (1985).

very simple requirement that entities be represented over many units—or, more generally, over some relatively extended portion of the resources available for representing.[4] Lacking any better term, I will describe representations that are spread out in this sense as *extended*. Clearly, a representation can only be extended by comparison with some normal, minimum or standard form, which can vary from case to case and style to style. Thus in typical connectionist networks the benchmark is one computing unit to every item, and relative to this a representation is extended if it uses many units for every item. In the brain the most plausible minimal unit is presumably the neuron (as in "grandmother" or "yellow Volkswagen" cells). Note however that in some other cases, such as optical holography—generally taken to be a paradigm example of distributed representation—there is no obvious parallel, because the surface of a photographic plate does not come naturally partitioned.

This bare notion of extendedness may seem trivial, but it is a very common theme in characterizations of distribution; indeed, on some occasions distribution is described solely in such terms.[5] It is therefore interesting to see what, if anything, is gained by distributing even in this minimal sense.

An important practical concern is worth mentioning first: extendedness can buy a certain kind of reliability or robustness. If an item is represented over many locations in such a way that no particular location is crucial to overall efficacy, then the system can withstand small and isolated damage or noise relatively well. This point is illustrated by the benefits of redundancy. Duplicating a given representation many times obviously increases the ability of the whole collection to convey the same content under adverse conditions. Thus, one reason for the industrious copying of medieval manuscripts was to ensure that if any one were lost, the same text would be preserved elsewhere—a point with a modern counterpart for users of word processors. An extended representation need not be simply redundant, however. Instead of activating a single neuron to represent a given perceptual item, the brain activates a vast number, forming an overall pattern for the same purpose, but where each neuron is tuned in a slightly different way to the retinal input. Loss of any particular neuron, or noise in the system, has almost no effect on the overall effectiveness of this representation,

[4] This is the point at which characterizations of distribution in the sense of interest to us here comes closest to the mainstream computer science conception. In this latter usage "distributed" applies to systems where storage or processing responsibilities are not restricted to any single computer but rather are spread through a connected network of machines; coordinating these various computers to achieve a single task then becomes a major design problem.

[5] Kosslyn and Hatfield (1984), for example, claim that: "In the brain, the best current guess is that information is not stored in a given location. Rather, information appears to be distributed across numerous locations" (p. 1030). Fodor & Pylyshyn (1988) think that "To claim that a node is distributed is presumably to claim that its states of activation correspond to patterns of neural activity—to aggregates of neural 'units'—rather than to activations of single neurons . . ." (p. 19). See also McClelland, Rumelhart, and Hinton (1986, p. 33); Papert (1988, p. 11; P. M. Churchland (1986, p. 289); Tienson (1987, pp. 10–11); Tye (1987, p. 170).

which is fortunate, given how noisy neurons are and the rate at which we lose them.

Whether this advantage of noise or damage resistance in fact accrues to a given case of extended representation depends very much on the form of encoding involved. A particular number N might be represented in a digital computer either in binary form or, in a more extended fashion, as a string of bits of length N; neither has any particular advantage of reliability over the other, even though the latter uses vastly more resources. This is just to stress again the point that extendedness must be achieved in such a way that no particular unit or location is crucial, a condition violated in both these cases. Further, the portion of the resources involved should be large not only relative to some theoretical minimum (such as the unit, neuron, or location), but also relative to the scale of likely damage or noise in the system itself.

Whether a representation is extended is independent of the stronger requirement that a given representation take the form of a distinctive pattern over that larger portion of the resources. Despite this independence, an important advantage in distributing representations in this sense is that it makes possible the use of a distinctive pattern for each distinct item. Indeed, such an approach will be essential if we want to represent a number of different items over the same set of units. Many authors, especially connectionists and commentators on connectionism, claim that the essence of distribution is to be found in this shift to the level of overall patterns, or, more generally, to characteristic overall states of the network or system. Rosenfeld and Touretzky (1988), for example, have defined schemes of distributed representation as those in which "each entity is represented by a pattern of activity over many units (p. 463).[6]

These patterns might be completely unrelated; they might be chosen at random, or it might suit one's computational purposes to choose patterns simply so as to maximize the distinctness of any two representations. On the other hand, an important reason for moving to characteristic patterns for the representing of each item is that the internal structures of these patterns can be systematically related, both to each other and to the nature of the items to be represented, thereby making the overall scheme more useful in certain ways. There are many ways to develop pattern-based schemes of representation in which the internal structures of the patterns have this kind of systematic semantic significance, but one in particular has been especially popular in connectionism. On this approach

[6] Consider also Feldman (1989): "The most compact representation possible would have a unique unit dedicated to each concept. If we assume that a unit corresponds to one neuron, then this is the grandmother cell or pontifical cell theory. The other extreme would have each concept represented as a pattern of activity in all the units in the system. This is well known as the holographic model of memory, and it is the most highly distributed theory that we will consider" (p. 71). Other such characterizations include Bechtel (1987, p. 19); Churchland and Sejnowski (1989, p. 30ff.); P. S. Churchland (1989, p. 118); Cummins (1989, p. 147); Smolensky (1987b, p. 144); Lloyd (1989, p. 110); Touretzky (1986, p. 523).

individual processing units pick out (micro)features, which are simply aspects of the domain, though usually at a much finer grain than that of the primary items to be represented. In a well known example, in order to represent a kind of room, we first assign to individual units features that are typically found in various kinds of rooms, such as *sofa*, *TV*, *ceiling and stove*.[7] Each different kind of room can then be represented by means of a distinctive pattern over these units—that is, that pattern that picks out all and only the relevant features. In this way it is possible to generate patterns for representing items in which semantic differences are built directly into the internal structure.

The popularity of this approach in connectionism has led to a common misconception of distributed representation as somehow essentially related to the notion of a microfeatural semantics for individual units. Thus, some influential commentators, including notably Pinker and Prince, have seized on the deployment of microfeatures as the crucial feature distinguishing genuinely distributed representations from other pattern-based schemes.[8] Meanwhile, others (such as Lloyd, 1989, p. 106) urged precisely the opposite—that is, that the distinguishing mark of genuinely distributed representation is that individual units do not have any semantic significance, microfeatural or otherwise. Though this disagreement is partly terminological, there is an important issue at stake here. As has been pointed out by connectionists in a number of places, it is quite possible to develop a scheme in which every item is represented by means of a semantically significant pattern over a set of units without individual units having any particular microfeatural significance at all.[9] The employment of microfeatures (such as "Wickelfeatures") is just one relatively convenient method for generat-

[7] This example is drawn from the Schema model (Rumelhart, Smolensky, McClelland, & Hinton, 1986).

[8] Pinker and Prince (1988) claim that "PDP models . . . rely on 'distributed' representations: a large scale entity is represented by a pattern of activation over a set of units rather than by turning on a single unit dedicated to it. This would be a strictly implementational claim, orthogonal to the differences between connectionist and symbol-processing theories, were it not for an additional aspect: the units have semantic content; they stand for (that is, they are turned on in response to) specific properties of the entity, and the entity is thus represented solely in terms of which of those properties it has" (p. 115). There is some precedent in connectionist writings for this position; thus Rumelhart, Hinton, and McClelland (1986) claim that "In some models these units may represent particular conceptual objects such as features, letters, words, or concepts; in others they are simply abstract elements over which meaningful patterns can be defined. When we speak of a distributed representation, we mean one in which the units represent small, feature-like entities. In this case it is the pattern as a whole that is the meaningful level of analysis. This should be contrasted to a one-unit-one-concept representational system in which single units represent entire concepts or other large meaningful entities" (p. 47). See also Clark (1989, p. 94).

[9] For example: "Another possibility . . . [is] that the knowledge about any individual pattern is not stored in the connections of a special unit reserved for that pattern, but is distributed over the connections among a large number of processing units. . . . The units in these collections may themselves correspond to conceptual primitives, or they may have no particular meaning as individuals" (McClelland, Rumelhart, & Hinton, 1986, p. 33).

ing such patterns, a method that can readily be discarded if it fails to provide a useful overall scheme of representation. Connectionist work is by no means deeply committed to the existence of microfeatural analyses of task domains, or to simple feature-based semantics for their representations (the failings of which are readily acknowledged on all sides), and the use of microfeatures is best regarded as essentially incidental to the distributedness of representations.

When one wants to model cognitive functions, it is particularly useful to have internal structure in the representations reflecting semantic properties, for differences in internal structure will have direct causal effects that can systematically influence the direction of processing. Thus some connectionists have argued that using patterns of activity over many units, rather than activation in a single unit, makes possible processes that, by virtue of their sensitivity to the internal structure of the representation themselves, are indirectly sensitive to the nature of the represented items.[10] Typically, choosing a strictly localist style of representation will preclude interesting internal structural differences between representations, and so, to compensate, the representations of each item will have to be supplemented by further knowledge about how it should be processed (stored, for example, in symbolic rules, or in the connectivity pattern among units). In one sense, of course, connectionists are here simply reiterating an old familiar point. Thus we use a compositional language rather than an unbounded set of primitives precisely because the internal structure of compositional representations has a systematic semantic and computational significance. If there is anything novel in the connectionist emphasis on using patterns with semantically significant internal structure, it lies in the highly controversial claim that the pattern-based schemes naturally implementable in neural networks are expressively more rich than any feasible compositional language. It is often claimed, for example, that by using patterns of neural activity as the fundamental mode of representation, connectionists are able to represent fine shades of meaning in a way that is fundamentally or at least practically impossible for normal symbolic schemes.

One way to generate representations with both kinds of advantages discussed so far (i.e., robustness and rich internal structure) is by coarse-coding. This places at least two special conditions on the kinds of features assigned to individual units. First, they must be coarse, where again this is always relative to some intuitive or perhaps theoretically defined standard or minimum. One way the notion of a coarse assignment is often expressed is in terms of the size of the receptive field of a unit, that is, those aspects or features of the domain with

[10] For example, Anderson and Hinton (1981, pp. 11–12); Hinton (1981). As Hinton puts it: "the 'direct content' of a concept (its set of microfeatures) interacts in interesting ways with its 'associative content' (its links to other concepts). The reason for this interaction, of course, is that the associative content is caused by the direct content . . ." (p. 175).

which activity in a unit is correlated. Coarse assignments are those that give individual units relatively wide receptive fields. Thus, individual neurons in the visual cortex are known to pick up on features present in the visual input such as lines at a certain orientation, but they are often coarse-coded in the sense that the band of orientations to which the neuron will actually react is quite wide. Coarse assignments are often disjunctive; for example, to generate coarse-coded representations of rooms (as in the above example) we could have individual connectionist units standing not for individual features such as sofa, but for disjunctions of features (e.g., sofa or TV). This principle carries over to the example of neurons in the visual system, since neurons generally have wide receptive fields along a number of different dimensions at once (e.g., orientation, location, movement).

The second condition is that the assignments be overlapping; if one unit picks out sofa or TV, the next should pick out TV or ceiling, and so on. Again, visual neurons generally satisfy this condition; if one neuron has as its receptive field a certain portion of the visual field, another will cover an overlapping portion, and so on. A representation of a whole item, such as a room or a visual scene, is then just a distinctive pattern over the full set of units. Such a representation has the advantage of robustness, because the large and overlapping nature of the receptive fields of the units pretty much ensures that no individual unit or group of units is crucial to successful representing. These representations also possess semantically significant internal structure, because the particular pattern used to represent an item is determined by the nature of that item, and so similarities and differences among the items to be represented will be directly reflected in similarities and differences among the representations themselves.

In numerous places in the connectionist literature distribution is seen as identical with, or at least deeply bound up with, coarse coding.[11] Coarse coded representations satisfy the first of the conditions in the standard connectionist definition quoted at the start of this section, because they are characteristic patterns of activity over groups of many units. Moreover, they also satisfy the second requirement, that each unit be involved in representing many different entities, because the representations of other items are simply different patterns of activity over the same group of units. In fact, individual units can even be involved in representing many different entities at the same time, for the characteristic patterns for two different entities can be activated at once over the same set of units. The representings of the two different entities can in this way be

[11] Touretzky and Hinton (1988): "We have rejected this idea in favor of a distributed or "coarse coded" representation . . . each particular face is almost certainly encoded as a pattern of activity distributed over quite a large number of units, each of which responds to a subset of the possible faces . . ." (p. 426–427). See also Sejnowski (1981, p. 191); P. S. Churchland (1986, p. 459); Horgan and Tienson (1987, p. 105).

superimposed on each other. One might initially suppose that this would introduce all kinds of deleterious interference effects, but—depending very much on the details of the coarse coding scheme in question and the entities being represented—it turns out to be possible to superimpose patterns while preserving the functional independence of the representations. An excellent example is the working memory in Touretzky and Hinton's (1988) Distributed Connectionist Production System.

The fact that representations can be superimposed in this way highlights a crucial ambiguity in the apparently simple notion of "localist" representation with which distribution is often contrasted. When a contrast with extended representation is in order, the relevant sense of "local" is, roughly, that of restriction in extent. When a contrast with superimposed representation is in order, however, the relevant sense of "local" is that of discrete, separated or nonoverlapping. These are quite different and indeed independent properties, and for a clear understanding of distribution it is crucial they be carefully distinguished. One author who does so very clearly is Murdock (1979):

> The idea of distributed storage does not necessarily imply that information is physically spread out over a large area. Rather, the key issue is whether memory traces, whatever their nature, are separate or combined. With localized storage, each trace is separate; so whether it is boxes or bins, there is one trace per location. Distributed memory takes the opposite position, where combined (superimposed) traces can be stored and there is no individual representation for a given item.[12] (p. 111)

He cannot however afford to discount entirely the importance of being spread over a large area, because distribution in this weaker sense turns out in practice to be a necessary condition for representations to be effectively superimposed. If individual memories were encoded in the brain by single neurons, it would be very unlikely that memories could be stored in a superimposed fashion (many memories to one neuron). It is only because each memory is stored across a wide network of neurons that it is possible, in practice, to store many memories over the same set of neurons.

The superimposition of representings, in some form or other, is probably the single most common theme in characterizations of distribution, especially when we look beyond the connectionist literature. In many characterizations it is clearly the dominant theme. According to McClelland and Rumelhart (1986b),

[12] Kohonen (1984) is also clear about the distinction: "the spatial distributedness of memory traces, the central characteristic of holograms, may mean either of the following two facts: (i) Elements in a data set are spread by a transformation over a memory area, but different data sets are always stored in separate areas. (ii) Several data sets are superimposed on the same medium in a distributed form" (p. 81).

for example, "[Distributed] models hold that all memories, old and new, are stored in the same set of connections . . ." (p. 504)[13].

Under what circumstances is it correct to say that two representings are superposed? First, some points of terminology: When two representations are effectively superposed, they become a single new item representing both the original contents. Thus we should avoid thinking of superposed representations, unless for some reason we deliberately want to keep the separate identities of the originals in mind. Rather, we should think of the superposed representings achieved by a single representation. Second, the terms *superimposed* and *superposed* are only largely synonymous. In what follows I will use *superposed* on the ground that it is shorter, although rejecting certain domain-specific connotations such as the notion of wave addition in physics. *Superposed* as I intend it is defined in what follows.

Intuitively, the representings of two distinct items are superposed if they are coextensive—if, in other words, they occupy the same portion of the resources available for representing. Thus in connectionist networks we can have different items stored as patterns of activity over the same set of units, or multiple different associations encoded in one set of weights. This point can be stated a little more precisely as follows. Suppose we have some accurate way to measure the amount of the resources involved in representing a given content item C.[14] Then a representation R of an item C is conservative if the amount of the resources involved in representing C is equal to R itself (no more and no less). A representation R of a series of items c_i is superposed just in case R is a conservative representation of each c_i.

This characterization is completely general, but can easily be fleshed out in many different ways. For example, in developing his tensor product scheme for representing structured items in connectionist networks, Smolensky (1987b) offers a formal definition of superposition in terms of vector addition. The result of adding two vectors is a new vector that, under the scheme in question, is taken to represent the same items as both the originals. Since the portion of the resources implicated in representing each item is now exactly coextensive—that is, just the whole new vector itself—the representings are superposed in exactly the sense just outlined.[15]

[13] Kohonen (1984) claims that "In distributed memories, every memory element or fragment of memory medium holds traces from many stored items, i.e., the representations are superimposed on each other. On the other hand, every piece of stored information is spread over a large area." (p. 11). For other examples see McClelland and Rumelhart (1986a, p. 176); Papert (1988, p. 12); Rosenberg and Sejnowski (1986), p. 75); Sejnowski (1981, p. 191).

[14] How exactly this amount is determined is not the concern here, but, for a first pass, suppose that it includes any portion of the representational resources such that variation in that portion is systematically correlated with variation in C, where "systematically correlated" is essentially relative to processes and relations privileged by the particular scheme of representation being employed.

[15] See Smolensky (1987a) s.2.2.2.

Sometimes it is appropriate to say both that a representation has just a single content, and that the representing of that content involves superposition. An example is an optical hologram of a single scene. In this case there is a single content, but, given the way holograms are constructed, each part of scene is represented over the whole surface. Such cases are naturally covered by the above characterization, for the items c_i that are each conservatively represented by R are now simply the parts of a single overall content C. Superposition is essentially the same, whether it applies to a set of items or to the parts of a single decomposable item. The only different is one of convenience, that is, where we find it natural to think of the primary level for identification of distinct items or contents.

Superposition appears to come in something like degrees in at least three ways. First, R was required above to be a conservative representation of each c_i, with the consequence that the resources involved in representing each item are identical. This is the most interesting case, but various weaker notions can be defined by relaxing this condition, allowing weaker overlapping relations between some or all pairs of representings. This accounts for the intuition that there is some kind of superposition occurring in coarse-coded schemes. In such schemes the receptive field of a given unit generally overlaps significantly with those of its neighbors. Consequently, particular features of the represented item will tend to be represented in overlapping (i.e., partially superposed) regions of the representation itself. However, we do not have full superposition in the representings of the individual various features; it is merely a neighborly overlapping effect.

Second, whether the amounts of resources involved in representing distinct items are the same depends very much on how we measure the resources, and in particular on our choice of dimensions along which to require that the amounts be identical. To take a very simple example: A family photo album captures many scenes, but not in a superposed fashion, for each distinct scene is encoded by a given photograph. This assumes that we are considering the album along the natural spatial dimensions. If we consider the album as extended through time, and ask how much of the album is involved in representing each scene, the answer will be roughly the same for every one, suggesting that the representings are superposed. We could take this as indicating that the degree of superposition in a given case depends on the nature or the number of dimensions along which coextensiveness of representings obtains. A better view is that such examples highlight the requirement, for genuine superposition, that representings be coextensive in all dimensions, or at least in all those dimensions where variation makes a real difference to the semantic significance of the representation.

Finally, a third way superposition comes in degrees is in how many contents or content parts are represented over the same space. It makes intuitive sense that the more distinct contents represented over the same space, the more strongly superposed the representation. Holograms are a good example, because the relevant scene-portions are extremely fine-grained, no larger than point sources; each of these very fine portions is represented over the whole surface of the hologram. The strongest possible notion of superposition, then, would require that R be a conservative representation of arbitrarily fine portions of the overall content. It is not difficult to construct artificial examples of this kind of extreme superposition. Suppose for example we were to treat a function as represented by its Fourier transform. Each point in the transform function is obtained by multiplying the whole function with a distinct trigonometric function and integrating over the result. Each point in the original function, then, is effectively represented over the whole transform; thus there is full superposition of arbitrarily fine slices of the original. This kind of extreme case is however not the norm; indeed, often the idea of arbitrarily fine divisions of the content makes little sense.

A metaphor often used to illustrate the difference between (discrete) localist schemes and superposed schemes is that of a filing cabinet. In the ideal filing cabinet every distinct item to be represented is encoded on a separate sheet of paper, and the sheets are then placed side by side in cabinet drawers. Because every item is stored separately, every item can be accessed independently of all the others, and the modification, removal or destruction of any one piece does not affect any others. If representation in the cabinet were fully superposed, by contrast, there would be no separate location for each discrete item; rather, the whole cabinet would be representing every item without any more fine-grained correspondence of sheets or locations to individual items. Accessing the representation of one item is, in an obvious sense, accessing the representation of all items; and modifying or destroying the representing of one cannot but affect the representing of all (though the effect is not necessarily harmful).

Superposed schemes thus differ fundamentally from more standard localist varieties. In general, schemes of representation define a space of allowable representations and set up a correspondence with the space of items or contents to be represented. We are accustomed to thinking of such schemes as setting up a roughly isomorphic correspondence—that is, there is a distinct representation for every item to be represented, and the structure of the space of representations systematically corresponds (in a way that can only be characterized relative to the scheme in question) to the structure of the space of possible contents. Thus, languages usually define an infinite array of distinct expression types, which are then put in correspondence with distinct items or states of affairs; and standard methods of generating images aim at finding a distinct image for every scene. When this kind of discrete correspondence fails to be the case—when, for example, the representational scheme assigns two distinct contents to the one representation (ambiguity), or two distinct representations to the one content (synonymy, redundancy)—we usually think of it as some kind of aberration or defect in the scheme, to be ironed out if we were to set about improving the scheme.

The notion of superposed representation overthrows this whole familiar picture, for superposition aims precisely at finding one point in the space of representations that can serve as the representation of multiple contents. For schemes in which R is a representation of c only if there is some non-arbitrary structural or functional relationship between R and c (if, for example, c itself or important information about c an be recovered from R) this becomes a formidable requirement, in the sense that actually finding such a point, one that can do double or multiple semantic duty, may be quite difficult. It is a nontrivial requirement on the way a scheme is set up that it be rich enough to be able to generate such representations. Consider the well-known example of a simple linear network incapable of computing the "XOR" function. The fundamental problem here is that no one point in the space of connection weights is capable of performing all the transformations from input to output which define that function. Representing these transformations in a fully superposed fashion requires moving to a different network structure, offering a different space of possible representations in the connections.

Suppose one's aim is to store a number of items. A localist scheme represents each item in its stored form exactly as it was represented previously; this is an essential part of the filing cabinet analogy. A superposed scheme, however, must find a single point in the space that can function properly as the representation of every item to be stored. Hence superposed storage involves transformation, and if an item is to be recovered in its original form, this transformation must be reversed. This brings us to another common theme in characterizations of distributed representation, the notion that although localist schemes store things as they are, distributed schemes must recreate the originals, or something like them, on demand: as Rumelhart and Norman (1981) stated, "Information is better thought of as 'evoked' than 'found' " (p. 3).[16]

It was pointed out just above that finding a point in the space of representations that can do multiple semantic duty is often difficult, and there is no guarantee that there is any one point that can perform its multiple duties perfectly or even at all. A good procedure for producing superposed representations will find a point that performs optimally in representing the full set of items, but such a point may still perform imperfectly with respect to a particular item. Various different schemes for producing superposed representations utilize different encoding processes of varying degrees of complexity and subtlety. It is typically the case, however, that when the representings of various different items are super-

[16] A good example is also found in McClelland, Rumelhart, and Hinton (1986): "In most models, knowledge is stored as a static copy of a pattern. Retrieval amounts to finding the pattern in long-term memory and copying it into a buffer or working memory. There is no real difference between the stored representation in long-term memory and the active representation in working memory. In PDP models, though, this is not the case. In these models, the patterns themselves are not stored. Rather, what is stored is the *connection strengths* between units that allow these patterns to be re-created" (p. 31). For a particularly strong version, compare Murdock (1982): "What is stored is not a 'wax tablet' or graven image; in fact, what is stored is not in any sense any sort of an item at all" (p. 623).

posed, these various representings influence and perhaps interfere with each other; for example, the superposed recording of two different scenes in one hologram results in partially degraded performance for each. In general, then, whether due to limitations inherent in the representational space, or limitations of the superposing process itself, superposed representations often exhibit imperfections that arise precisely because various representings are being superposed. This, then, is another recurrent feature in descriptions of distribution: "Memory systems in which distinct stored items are spread over the storage elements and in which items can mix with each other are often referred to as 'distributed' or 'holographic' " (Anderson, 1977, p. 30).[17] Although this might pose a technical problem for anyone interested in devising efficient distributed storage methods, it might also be a crucial feature of systems utilizing distributed representations in psychological modeling, because these imperfect performance characteristics may turn out to be useful in explaining certain features of human cognitive performance.

There is one particularly important manner in which this mixing or interference is often manifested. Suppose we have a number of items to be represented, each of which is a variant on a theme (or perhaps a few central themes). Suppose, for example, we wish to store a series of vectors exhibiting a certain structural tendency or pattern in common, though there is significant random variation from one to the next. Then when the representings of all the items are superposed, those respects in which they are similar will naturally be reinforced, and those in which they differ will tend to be in conflict and hence cancel each other out. Though this effect depends very much on both the details of the particular encoding process that generates the superposed representation and the particular items to be stored, what can often happen is that the resulting representation performs well insofar as it is representing the common features of each item, but poorly insofar as it is representing unique variations. The representation, in other words, has generalized; it has emphasized the central tendencies from among a set of exemplars, and now disregards or downplays particular differences.[18] This generalization can be advantageous at a later point, when the representation is evaluated with regard to its performance as a representation of some item it was not originally designed to handle. In such a case, the representation will perform acceptably just insofar as that item can be treated as exemplifying the central tendencies previously extracted.

In any localist scheme, each distinct item is represented by means of a proprietary chunk of the available resources. Because resources are usually finite,

[17] The same theme is taken up by Rumelhart and McClelland (1986): "Associations are simply stored in the network, but because we have a *superpositional* memory, similar patterns blend into one another and reinforce each other" (p. 267).

[18] Of course, if it is desirable to actually *preserve* these individual differences, encoding schemes can be designed and implemented accordingly.

this puts a strict upper limit on the number of items that can be represented, and attempts to exceed this limit can have more or less unfortunate consequences. Superposed schemes, by contrast, attempt to find a single state of one (possibly large) chunk of the resources that functions as a representation of all the items to be stored. Storing another item does not take up more resources; rather, it involves transforming the state of the same resources. The issue, then, is not whether there are resources available for storing the item at all; it is rather how well the new state functions as a representation of all the stored items. Depending on the particular scheme in question, it is typically the case that storing more items results in gradual worsening of performance across all or most items rather than an inability to store any particular one. This is the much vaunted *graceful degradation* of distributed systems, and it clearly follows from the nature of superposed representation (or rather, one important aspect of it).[19]

EQUIPOTENTIALITY

I have stressed the fundamental gulf between superposed schemes of representation and more familiar localist schemes, but distribution has on occasion been associated with even more radical possibilities. Consider Lashley's (1950) famous claim about the representation in the brain: "It is not possible to demonstrate the isolated localization of a memory trace anywhere within the nervous system. Limited regions may be essential for learning or retention of a particular activity, but within such regions the parts are functionally equivalent. The engram is represented throughout the area. . ." (p. 477). The first sentence appears to be a clear statement of superposition. The second sentence however claims that all memory traces are contained in all parts of the brain region. This is a much stronger claim, for it is easy enough to envisage a case of fully superposed representings, but where parts of the overall representation do not have the same content as the whole. It is clear that Lashley was aware of the difference between these two properties, and in general claimed that the stronger one was true of the brain. Thus, in earlier work, he had formulated the famous principle of neural equipotentiality, "the apparent capacity of any intact part of a functional area to carry out, with or without reduction in efficiency, the functions which are lost by destruction of the whole. . ." (Lashley, 1929, p. 25). For current purposes we should modify this principle and think of a representation as equipotential just in case each part of that representation has the same semantic significance as the whole. Every part represents just what the whole represents.

Some have explicitly maintained that equipotentiality is the essence of dis-

[19] Note that the term *graceful degradation*, like the term *distributed representation*, has many senses; the sense described here is particularly appropriate to a narrow focus on the issue of representation.

tributed representation,[20] and this idea is implicit in much other discussion. Thus, insofar as Lashley is regarded as having determined that memories are represented in distributed fashion in the brain—a very common move—distribution is being at least implicitly equated with equipotentiality. The same is true when we regard optical holograms as paradigms of distributed representation, for equipotentiality is one of the most well-known features of at least certain varieties of hologram.[21] It is therefore worth spending some time clarifying this notion of equipotentiality, especially in its relation to superposition.

Equipotentiality requires that the various parts of a representation R have the same content as R itself. There is an obvious symmetry here with the notion of superposition, which required that the various parts of the overall content have the same representing. This suggests the symmetrically opposed definition: R is an equipotential representation of C just in case every part r_i of R is a representation of C.[22] Unfortunately, this form of the definition seems to allow for tedious counterexamples involving the simple duplication or replication of a given self-sufficient representation. An example is Venetian wallpaper, which duplicates hundreds of identical discrete little sketches of the Rialto all across the wall. Every sketch is representing the Rialto (or Venice), and the whole wall certainly represents nothing more, and so it would seem that the wallpaper is equipotential. Surely equipotentiality is a more interesting phenomenon than this!

One problem with these merely redundant representation is that their equipotentiality, such as it is, stops at a certain fixed level: portions of the overall representation smaller than an individual Rialto sketch do not have the same content as the whole. One way to rule out such cases, then, would be to require that arbitrary portions of R have the same content as the whole. Thus, R is equipotential with respect to C just in case every portion r_i of R is a representation of C for every division of R into parts; that is, no matter how you slice it, each part still has the same content as the whole. This suggests an asymmetry between superposition and equipotentiality: because although superposition is an interesting phenomenon for any given discrete division of the content into more than one part, equipotentiality only appears to be interesting for arbitrary divisions.

It is doubtful, however, whether equipotentiality in this sense is ever attained.

[20] Westlake (1970): ". . . the property of distributedness, which is displayed only by holographic processes. This property, an attribute of certain types of holograms, permits any small portion of the hologram to reconstruct the entire original scene recorded by the hologram . . ." (p. 129). See also Pribram (1969, p. 75).

[21] Leith and Uptanicks (1965): "each part of the hologram, no matter how small, can reproduce the entire image; thus the hologram can be broken into small fragments, each of which can be used to construct a complete image" (p. 31).

[22] We cannot require that eachr$_i$ be a *conservative* representation of C, for if r_i is conservative, then r_i would have to include all of the resources involved in representing C—that is, R could only have only been divided into one part. This trivialises equipotentiality.

For one thing, if we take fine enough portions of any representation it is unlikely we will have any content at all, let alone the full original content. Though holograms, for example, are taken to be equipotential, small enough portions fail to encode anything; all we need do is take some portion smaller than the wavelength of the illuminating beam and we cannot possibly recover anything of the original image. On reflection it would be truly remarkable for a nontrivial representation to have the same content in any portion, no matter how small, as is conveyed by the whole. Second, even reasonably large portions of the original representation typically are not identical in content with the original. Lashley himself was careful to qualify his principle of neural equipotentiality with a corresponding principle of mass action, which acknowledged that the smaller the chunk of brain remaining, the worse the performance. Each portion, it would seem, could not have had exactly the same content as the whole; for if it had, it would presumably have been able to generate the same performance. Similarly for the hologram: The whole scene is recoverable from each part, but only with proportionately reduced perspective and image quality.

Rescuing the notion of equipotentiality in the face of these objections requires a two-pronged strategy. On one hand, we need only require that all portions of some sufficient size (an inherently vague notion) represent the same as the whole. Mere redundancy then becomes one trivial way of achieving this effect, practical in some contexts, such as the preservation of medieval wisdom, but manifestly implausible in others, such as encoding memories in the brain. More interesting methods, such as the convolution transformation underlying holography, vary parts of R smoothly and systematically as a function of the whole content. Second, we need to explain some sense in which these sufficiently large portions represent the same as the whole despite slight variation or degradation. Intuitively, the portions are all still in some important sense representing the same thing, regardless of the degradation in performance. Sustaining this intuition means formulating some kind of distinction between what a representation is of and how good it is as a representation of that content; or, in other words, a kind of semantic character versus quality distinction. A representation would then be equipotential insofar as all sufficiently large portions have, not the same "content," but rather the same semantic character as the whole, even if at lower quality.

When do two representations have the same semantic character in the relevant sense? We need here some notion of a privileged or important dimension of the content, such that each part can be seen as semantically coextensive along that dimension, although perhaps varying on others. To illustrate: Why do a hologram and its portion have the same semantic character, even if the portion performs significantly worse in generating the whole scene? The answer is that crudely spatial dimensions of the encoded scene are accorded a certain kind of priority, and what the portion does is recreate the whole scene along these dimensions, albeit in a degraded way. In short, two representations have the same character if their performance is essentially equivalent along what happens to be the intuitively important dimension, though it may vary on others. Lashley regarded regions of the rat brain as equipotential because, with only portions of the region remaining, a rat could perform the same tasks, even if more slowly or clumsily. Hence the important dimension here is simply the fact of performance; speed or agility is relegated to a lesser importance and so is just a matter of quality, not character.

This description of the character versus quality distinction is a start, but it is by no means complete because it does not give any general guidelines for determining what the important dimension is in a particular case. It is unlikely, however, that there could be any such general guidelines, because the relevant dimension changes from case to case according to interests and purposes that can vary greatly in ways that have little or nothing to do with intrinsic properties of the representations themselves. It is natural to be impressed by the fact that each portion of the hologram reproduces the same spatial extent of the image. Yet consider the case of an aerospace engineer devising a holographic display unit for a jet fighter. Here the primary advantage a holographic display has over a regular screen is that the generated image conveys depth effects at sufficiently high resolution to assist the pilot in operating the plane. Both depth effects and resolution would presumably be lost, however, if only a portion of a hologram were employed; hence, from the engineer's point of view, the hologram is not at all equipotential. Though the hologram itself remains the same, from one naive point of view it counts as equipotential, although from another point of view it does not. Because the general character versus quality distinction depends on the notion of a privileged dimension, that distinction itself is rendered inherently flexible, even vague; and consequently the very idea of equipotentiality is without any very firm foundation.

It has already been pointed out that superposition does not entail equipotentiality. Given the symmetry between the concepts, it should not be surprising that the converse is also true: Equipotentiality does not, in general, entail superposition. Thus, consider again the trivial case of Venetian wallpaper. If we consider only parts larger than a certain minimum size, then each part has the same content as the whole. Yet there is clearly one scattered portion of the wall where the Rialto is found multiply depicted, and likewise one portion where the gondola is found, and these portions are entirely discrete, as can be seen by the fact that we could paint over all Rialto depictions while leaving all gondola depictions intact. The wallpaper is therefore trivially equipotential with respect to the whole scene, but not superposed with respect to the Rialto and the gondola.

The relation between these concepts is however more intimate than these independence claims suggest. Notice for example that full equipotentiality (i.e., equipotentiality of arbitrarily fine portions of R) immediately entails full superposition. If the whole content is encoded in every part of the representation, no matter how fine, it follows that every part of the content must be encoded over the whole representation.[23] Further, it turns out that standard methods for generating real instances of equipotential representation do in fact vary every part of the representation as a function of the whole content, thereby guaranteeing superposition. Conversely, common methods for developing superposed representations often produce something akin to equipotentiality as a side-effect. In connectionism, for example, it is common to represent transformations from input to output in the one set of weights. It is a well-known feature of such representations that they are relatively impervious to localized damage or noise; thus, removing units or connections makes relatively little difference to overall performance (Wood, 1978). Another way to describe this situation is in terms of the equipotentiality of the representation: large enough portions effectively represent the same as the whole.

In general, insofar as there is equipotentiality, any portion of a representation can take over the tasks of the whole; in other words, equipotential representations are robust by their very nature. This brings the discussion of themes associated with distribution around a full circle. Robustness was seen to be a desirable consequence of at least some forms of merely extensive representation, but it dropped out of consideration in the discussion of superposition. Although superposed representations are often relatively robust, nothing in the definition of superposition itself guarantees this: although a series of items are represented over the same resources, it might be that all those resources are required for the effective representing of any one item. However, with equipotentiality, which is intuitively the strongest form of distribution of all, robustness is guaranteed.

Distributed systems or representations are often described as *holistic*. This is an extraordinarily vague term, and usually contributes nothing to our understanding of the phenomenon in question; nevertheless, with the above discussion in mind it is possible to sort out some things that might be intended. For example, describing a distributed representation as holistic might be a reference to the fact that, when a representation R is superposed, each part of R is involved in representing a number of items at once, and in that sense reflects the "whole" content. Similarly, in superposed schemes R functions as a representation of a number of items at once; in that sense, one state represents the whole content, or each item is only represented in the context of the whole content. Alternatively, describing distributed representations as holistic might be a reference to equipotentiality, where each part represents the "whole" content. Each of these senses gestures in the direction of some important aspect of distributed representation;

[23] The converse, however, does not in general hold: full superposition does not entail full equipotentiality. This can be seen from the Fourier transform example. Though every point in the original function is represented over the whole transform, it is not possible to recreate the whole original function from any given point in the transform. This difference is due to an asymmetry in the concepts themselves. Equipotentiality requires that each portion of R actually represent the whole content, although superposition requires only that each portion of R be involved in representing each part of the content.

however, superimposing them in the one (dare I say, holistic) concept results, in this case, in little more than a blur.[24]

WHAT IS DISTRIBUTION?

This discussion sketches just the broadest outlines of the current concept of distribution. It reveals something of the multiplicity of themes in the vicinity of distribution, and some of the extraordinary differences among characterizations previously put forward. But how does it bear on the wider project of finding some general kind of representation on the basis of which a genuine alternative to CTM might be constructed? What, after this discussion, can we say that distributed representation actually is?

In view of the multifarious nature of typical instances, and the wide variety of properties thought to be central to distribution, it may be tempting to suppose that distributed representation is really just some kind of "family resemblance" concept, gathering together in a loose way a heterogeneous collection of styles of representation that actually turn out to have no interesting properties in common. Such concepts are bad news for theorizing because they effectively preclude one from making interesting general claims about all members of that type. Consequently if a survey of current usage revealed that distribution was in fact such a concept, the speculative project of investigating alternatives to CTM constructed around a putative category of distributed representation would be stopped in its tracks.

This prospect is not a serious concern, however. Current usage has only a limited claim on our allegiance. The appropriate response, in the interests of conceptual clarification and scientific progress, would be simply to redraw the conceptual boundaries, thereby revising the current muddled use of the term. We would be providing, in other words, an explication of the concept: a new, precise account of distribution, based on the old confused version but displacing it. In the current case, moreover, there is a way of explicating the concept of distribution that is not excessively disruptive of current usage. It turns out that one theme—the superposition of representings—is both common to a large proportion of standard characterizations and true of most cases that intuition counts as paradigm instances. This opens the door to a proposal, still informal and far from precise, but nevertheless substantial: distribution is the superposition of representings, and distributed representations are those which belong to schemes defined around a core method of generating superposed representations.

One reason for supposing that superposition is the heart of distribution is that

[24] The alleged "holism" of distributed representation has nothing to do with the term *hologram.* Denis Gabor coined the term, from the Greek for *whole writing*, to bring out the fact that a hologram records all the information in a given wave of light (i.e., intensity and phase).

the various other themes and properties discussed above are either implausibly weak or implausibly strong. Thus, extensiveness alone cannot be what is important about distributed representation, because a wide variety of representations that are obviously not distributed in any interesting sense consume more than some theoretical minimum by way of representational resources. To count as genuinely distributed a representation must be more than simply spread over a large area, whether that be on the page, in the sky, in a computer memory or in the brain. Similarly, a distributed representation must also be more than just a "pattern" over some extended area, for virtually any representation will count as a pattern of some kind. This notion must be tightened somehow, yet various common approaches—such as the requirement that the patterns result from coarse-coding—fail to capture the natural class of distributed representations, for they end up excluding various standard cases to which such an approach is inapplicable. Coming from the other direction, equipotentiality cannot be regarded as a definitive feature of distribution. For one thing, it is not always clear that equipotentiality is a well-defined property: it all depends on whether we can privilege some dimension along which it makes sense to say that the full content is preserved in each portion of the representation. Equipotentiality also suffers the problem of exclusivity; all kinds of representations that appear to be distributed in an important sense fail to exhibit significant equipotentiality.

Superposition has neither of these failings. It is strong enough that very many kinds of representations do not count as superposed, yet it manages to subsume virtually all paradigm cases of distribution, whether these are drawn from the brain, connectionism, psychology, or optics. Moreover, superposition is a satisfying choice in other ways. As a structural feature it seems to be in deep-seated opposition to the standard kinds of localist schemes with which we have long been familiar. As pointed out above, usual forms of representation are designed so that, roughly, the structure of the domain of representations mirrors that of world itself; every different item to be represented is mapped to its own distinctive point or points in the space of representations. Superposed schemes violate this neat order, and do so inherently; thus the distinction between localist and superposed schemes marks a fundamental gulf in kinds of representation, a gulf suggesting that semantic superposition is not some incidental property that merely cross-classifies other forms of representation, but rather is the kind of property that manages to pick out a whole distinct genus of its own.

Much remains to be done if this speculation is to be firmly grounded. The notion of semantic superposition has to be precisely defined, and it must be shown that schemes of distributed representation can be effectively generated in such terms; this includes showing that a sufficient number of the intuitive paradigm cases of distribution are indeed characterizable as falling under superposed schemes. It would also have to be demonstrated that the loose claims of incompatibility made here between superposition and other supposedly localist styles, such as generically symbolic representation, stand up under closer scrutiny. If such elaboration were successful, however, we would be able—indeed, obliged—to investigate the possibility raised at the beginning of this paper: namely, that a theory of cognition might be constructed on a distributed foundation, a theory that would automatically count as an alternative to CTM.

Thorough investigation of this possibility would involve determining whether the generic category of distributed representation, defined in terms of superposition, satisfies the demanding criteria set forth earlier. It is entirely unclear at this stage, for example, whether distributed representations (of whatever particular variety) would be sufficiently powerful to represent effectively the kinds of information that underlie human cognitive performance. Much connectionist work in psychological modeling can be regarded as an empirical investigation of precisely this issue, but it is still much to early to expect any conclusive verdict, especially when we realize that those investigations are being carried out in the absence of any well-developed theory of the nature of distributed representation.

On the other hand, there are some reasons to be optimistic. I have suggested already that, from the preliminary perspective afforded by this overview, it is plausible to suppose that distributed representation is both appropriately general as a category and inherently non-symbolic, features that surely constitute a promising start. Moreover, distributed representations possess the desired deep affiliation with neural networks. Briefly, on the current proposal, a distributed representation is one in which each component of the representation is implicated in the representing of many items at once. The high degree of interconnectedness between processing units in neural networks constitutes excellent conditions for implementing this kind of dependence. This is not to say, of course, that any representation in a neural network must be distributed; on the contrary, it is manifestly possible to impose localist or even symbolic structures on neural mechanisms. It is to say that neural networks provide a very natural medium for implementing distributed representations; or rather, to put the point even more strongly: to insist on utilizing nondistributed representations in a neural network framework would be to stubbornly avoid capitalizing on some of the most important benefits of neural machinery. It would be akin to using digital electronic circuitry while stubbornly refusing to implement general purpose symbol processing.

This deep affiliation makes distributed representation attractive as a possible alternative to symbolic representation. There is, for example, nothing implausible in supposing that there are distributed representations to be found in the brain. Quite the contrary: Distribution is an empirically well-established feature of the neurological substrate in which our cognitive capabilities find their ultimate realization. Indeed, the biological reality of distributed representation is a principle so secure that it cannot even be counted as a discovery, for the concept itself first arose in attempts to describe the unusual kinds of representations found in the brain, and when it was proposed above that there may be a broad natural category of distributed representation, neural representations were taken to be paradigm instances. This intimate association with the actual machinery underlying human cognition stands in plain contrast with the biological remoteness of symbolic representations. Though CTM demands a language of thought, and CTM advocates insist that the expressions of this language are realized in the neural substrate, and consequently predict the eventual discovery of "symbols amongst the neurons," neuroscience has never yet stumbled across syntactically structured representations in the brain. This discrepancy only becomes more embarrassing to CTM as the sum of neuroscientific knowledge increases, and provides at least a prima facie argument in favor of any biologically motivated alternative.

Nevertheless, this proposal does not fall afoul of another quite different methodological constraint based on the supposed autonomy of psychology and neuroscience. CTM orthodoxy maintains that psychological generalizations are only to be found at a certain level of description of a system—roughly, the level at which the system's operation is most usefully understood as cognizing, that is, involving the transformation of representations. When CTM proposes symbolic representation as the form in which information must be stored and processed, symbolic representation is described in a sufficiently general way in which only the abstract structure is important. All kinds of implementational details become irrelevant, and, consequently, systems differing widely in their physical instantiation can be regarded as falling under the same psychological principles. To descend to the level of the system's particular implementation would be to descend to a level from which the true psychological generalizations are no longer visible. An immediate consequence of this view is that theories and models of cognitive functioning are held to be relevantly similar, or seen as falling under the one approach or paradigm, only insofar as they are similar at this relatively abstract level. Thus it becomes possible to maintain that the sole thread binding all CTM (or classical) approaches together is a commitment to symbolic representation together with operations that are sensitive to syntactic structure.

Recently however, with rapid advances in neuroscience and the resurgence of connectionism, it has become popular to maintain that cognitive functioning is not independent of, and can not be understood independently of, the details of the particular way in which cognition happens to be instantiated. Thus, in the case of human cognitive abilities, it is maintained that we must shift the focus of attention from abstract, purely psychological or "top down" investigations to the messy details of the actual neurobiological mechanisms themselves, their evolutionary context, and their very specific capabilities. Meanwhile it is maintained that the crucial feature bringing the wide diversity of approaches within the groundswell of opposition to CTM under a single paradigm is not to be found at the level of a form of representation but rather in terms of a general commitment to neurobiological authenticity.

An articulated concept of distributed representation offers a sound theoretical basis for reconciling these two apparently conflicting strains of thought. It possesses an inherent neurobiological plausibility, but without sacrificing the generality required of a genuinely psychological hypothesis. I have urged that distribution be understood in terms of the superposition of representings, a notion that is completely independent of details of the particular ways in which the superposition might be achieved. The wide variety of instances of distribution already known to exist—and effectively describable in terms of superposition—attests to the breadth of this characterization. Given a well-developed conception of distributed representation, it is possible to determine which kinds of operations are suited to processing distributed representations, and consequently what kinds of distributed mechanisms might subserve various different cognitive functions, without ever needing to descend to descriptions of particular hardware implementations. Arguments for or against such theories can then be formulated at this relatively abstract psychological level. It is in this light that we should understand many connectionist models of cognitive functions that, despite being based on networks of neural units, are highly remote from biological details. This approach can also be seen in the work of psychologists such as Metcalfe, who has proposed distributed mechanisms to account for various memory phenomena quite irrespective of how those mechanisms are in fact instantiated in the head; these proposals are then tested on the basis of straightforwardly psychological experiments. (For examples of Metcalfe's work see Eich, 1982; Metcalfe, 1989; also Murdock, 1979, 1982.) With high level descriptions of distributed mechanisms in hand, it is possible to continue on to construct much more specific models which specify, to some relevant level of detail, how these distributed mechanisms happen to be built up out of human wetware, a process which not only tests the hypotheses themselves but stimulates future developments.

In short, the concept of distributed representation enables us to preserve the insight that a theory of cognition is more than just a theory of how particular mechanisms perform their specific functions; but also, by virtue of its intimate relationship with connectionism and the brain, acknowledges the importance of detailed studies of the actual machinery underlying human cognitive performance. There need be no tension between studying cognition and studying particular neurobiological instantiations of cognition, because the general concept of distributed representation functions as the unifying principle. For this reason, there is also no need to insist that the central feature binding together a wide group of neuroscientific and connectionist alternatives to CTM is a concern with neurobiological plausibility, for we can now see the possibility of a deeper similarity between these various approaches, one that obtains at the level at which they count as theories of cognition.

REFERENCES

Anderson, J. A. (1977). Neural models with cognitive implications. In: D. Laberge & S. J. Samuels (Eds.), *Basic processes in reading* (pp. 27–90). Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. A., & Hinton, G. E. (1981). Models of information processing in the brain. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 9–48).

Bechtel, W. (1987). Connectionism and the philosophy of mind: An overview. *Southern Journal of Philosophy, 26* (Suppl.), 17–42.

Churchland, P. M. (1986). Some reductive strategies in cognitive neurobiology. *Mind, 95*, 279–309.

Churchland, P. S. (1986). *Neurophilosophy*. Cambridge, MA: MIT Press.

Churchland, P. S. (1989). From Descartes to neural networks. *Scientific American, 261*(1), 118.

Churchland, P. S., & Sejnowski, T. J. (1989). Neural representation and neural computation. In N. Nadel, L. A. Cooper, P. Culicover, & R. M. Harnish (Eds.), *Neural connections, mental computation* (pp. 15–48). Cambridge, MA: MIT Press.

Clark, A. (1989). *Microcognition: Philosophy, cognitive science and parallel distributed processing*. Cambridge, MA: Bradford Books/MIT Press.

Cummins, R. (1989). *Meaning and mental representation*. Harvard, MA: MIT Press.

Eich, J. M. (1982). A composite holographic recall memory. *Psychological Review, 89*, 627.

Feldman, J. A. (1989). Neural representation of conceptual knowledge. In L. Nadel, L. A. Cooper, P. Culicover & R. M. Harnish (Eds.), *Neural connections, mental computation* (pp. 69–103). Cambridge, MA: Bradford/MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 161–187). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: Foundations (pp. 77–109). Cambridge, MA: MIT Press.

Horgan, T., & Tienson, J. (1987). Settling into a new paradigm. *Southern Journal of Philosophy, 26*, (Suppl.), 97–113.

Kohonen, T. (1984). *Self-organization and associative memory*. New York: Springer-Verlag.

Kosslyn, S. M., & Hatfield, G. (1984). Representation without symbol systems. *Social Research, 51*, 1019–1045.

Lashley, K. S. (1929). *Brain mechanisms and intelligence*. Chicago: University of Chicago Press.

Leith, E. N., Uptanieks, J. (1965). Photography by Laser. *Scientific American, 212*, 24–35.

Lloyd, D. (1989). *Simple Minds*. Cambridge, MA: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1986a). A distributed model of human learning and memory. In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models* (pp. 170–215). Cambridge, MA: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1986b). Amnesia and distributed memory. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel distributed processing: explorations in the microstructure of cognition. Vol. 2: Psychological and biological models* (pp. 503–527). Cambridge, Mass.: MIT Press.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel distributed processing: explorations in the microstructure of cognition. Vol. 1: Foundations* (pp. 3–34). Cambridge, Mass.: MIT Press.

McEliece, R. J. (1985). The reliability of computer memories. *Scientific American, 252*(1), 88–95.

Metcalfe, J. (1989). Composite Holographic Associative Recall Model (CHARM) and Blended Memories in Eyewitness Testimony. *Proceedings of the 11th Annual Conference on the Cognitive Science Society* (pp. 307–314).

Murdock, B. B. (1979). Convolution and correlation in perception and memory. In L. G. Nilsson

(Ed.), *Perspectives on memory research: Essays in honor of Uppsala University's 500th Anniversary* (pp. 609–626). Hillsdale, NJ: Lawrence Erlbaum Associates.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*(6), 609–626.

Papert, S. (1988). One AI or many? *Daedalus, 117*(1), 1–14.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28,* 73–193.

Pribram, K. H. (1969). The neurophysiology of remembering. *Scientific American, 220,* 75.

Rosenberg, C. R., & Sejnowski, T. J. (1986). The spacing effect on NETtalk, a massively-parallel network. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 72–88).

Rosenfeld, R., & Touretzky, D. S. (1988). Coarse-coded symbol memories and their properties. *Complex Systems, 2,* 463–484.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In PDP1 (pp. 45–76).

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel distributed processing: explorations in the microstructure of cognition. Vol. 2: Psychological and biological models* (pp. 216–271). Cambridge, Mass.: MIT Press.

Rumelhart, D. E., & Norman, D. A. (1981). Introduction. In *Parallel Models of Associative Memory* (pp. 1–8). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel distributed processing: explorations in the microstructure of cognition. Vol. 2: Psychological and biological models* (pp. 7–57). Cambridge, Mass.: MIT Press.

Sejnowski, T. J. (1981). Skeleton filters in the brain. In G. E. Hinton & A. J. Anderson (Eds.), *Parallel models of associative memory* (pp. 189–212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Smolensky, P. (1987a). On variable binding and the representation of symbolic structure in connectionist systems. (Tech. Rep. CU-CS-355-87). Department of Computer Science, University of Colorado.

Smolensky, P. (1987b). The constituent structure of mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, 26 (Suppl.), 137–160.

Tienson, J. (1987). Introduction to connectionism. *Southern Journal of Philosophy, 26* (Suppl.), 1–16.

Touretzky, D. S. (1986). BoltzCONS: Reconciling connectionism with the recursive nature of stacks and trees. *Proceedings of the Eight Annual Conference of the Cognitive Science Society* (pp. 265–273).

Touretzky, D. S., & Hinton, G. E. (1988). A distributed connectionist production system. *Cognitive Science 12*, 423–466.

Tye, M. (1987). Representation in Pictorialism and Connectionism. *Southern Journal of Philosophy, 26* (Suppl.), 163–184.

Walters, D. (1987). Properties of connectionist variable representations. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 265–273).

Westlake, P. R. (1970). The possibilities of neural holographic processes within the brain. *Kybernetik, 7,* 129–153.

Wood, C. C. (1978). Variations on a theme by Lashley: Lesion experiments on the neural model of Anderson, Silverstein, Ritz and Jones. *Psychological Review, 85,* 582–591.