Modern Psychological Studies

Volume 3 | Number 1 Article 4

1995

An overview of parallel distributed processing

Jonathan P. Heyl
East Carolina University

An Overview of Parallel Distributed Processing

Jonathan P. Heyl

East Carolina University

Abstract

Parallel Distributed Processing (PDP), or Connectionism, is a frontier cognitive theory that is currently garnering considerable attention from a variety of fields. Briefly summarized herein are the theoretical foundations of the theory, the key elements observed in creating simulation computer programs, examples of its applications, and some comparisons with other models of cognition. A majority of the information is culled from Rumelhart and McClelland's (1986) two-volume introduction to the theory, while some concerns from the field and the theorists' accompanying responses are taken from a 1990 article by Hanson and Burr.

The theory of Parallel Distributed Processing (PDP), or "Connectionism," is currently enjoying some degree of attention, in part due to its fairly radical departure from the mainstream ideas about cognition that have dominated the field for the past several decades. The reactions to it range from unconditional acceptance to denial, but the attitude of the majority seems to reflect a "curious and interested but not entirely convinced" viewpoint. The fact that such an apparently large number of people in the field are displaying an interest tends to lend the theory a degree of legitimacy, or at least worthiness of further exploration and study, and is the view primarily reflected in this paper.

The majority of the information used to describe the theory in this paper is culled from Rumelhart and McClelland's two-volume explanation of PDP published in 1986, with the intent being to summarize and possibly simplify the major structural points. The latter part of this paper reviews some objections and concerns about the theory that followed a more recent article on PDP by Hanson and Burr (1990).

PDP is a difficult theory to summarize for several reasons. First, it is extremely complex mathematically, with numerous symbolic formulas to represent various functions. Secondly, its domain is broad, operating not only under the conventions of the field of psychology (diverse in itself), but expanding to encompass the fields of computer science, philosophy, and artificial intelligence. Thirdly, the theory is undergoing rapid refinement under a surge of experimentation and its subsequent new data, leaving some components of its original form approaching obsolescence.

"Parallel" processing denotes the simultaneous activity occurring in the brain at any one time, as opposed to processing of a serial nature. The idea is that such a vast amount of information is being processed so quickly in the performance of even the simplest tasks that time constraints rule out the possibility of exclusively serial operation. This is not necessarily meant to deny the existence of any serial processes, for it seems obvious that there are some (e.g., problem solving). Yet each step of these higher-order serial processes is made possible by, and is the result of, large-scale parallel processing.

Supporting the idea of parallel processing is the speed of the neuron, which is relatively slow (compared to a purely electronic device, and more specifically, a computer). Rumelhart and McClelland (1986) point out that "neurons operate in the time scale of milliseconds whereas computer components operate in the time scale of nanoseconds—a factor of 10^6 faster." Thus, observed human operations taking approximately one second to compute would be limited to performing only about one hundred sequential steps—an improbably low number given the complexity of functions like perception, speech analysis, etc. Also implying parallelism is the vast number of neurons we possess, estimated to be between 10^{10} and 10^{11}, each being an "active processing unit" and capable of

receiving input from many other neurons by means of simple excitatory or inhibitory impulses. These interconnections between neurons are usually short, as well as symmetrical, implying at least the possibility of back propagation (feed-backward as well as feed-forward impulses) and interactive activation (Rumelhart & McClelland, 1986).

The word "distributed" in the theory's title refers to the authors' contention that representations are not complete entities stored as a whole at some location in the brain. Rather, representations are distributed across some number of units (neurons) and are held in the "connective weights," or strengths, between them. When certain neurons are activated, their activation is fed through these weights to produce a pattern of activation representative of the "stored" item. When the memory is not activated, it resides only in connection strengths between neurons.

Part of the evidence supporting distributed storage is based on observations of how performance declines following brain damage. In many ways it reflects Lashley's (1929) Mass Action findings, wherein removal of brain tissue leads to "graceful degradation" of performance, rather than the complete loss of some motor function or the loss of an entire "concept." This is not intended to mean that there are no localized areas within the brain, but it does maintain that knowledge *within* these localized areas is distributed. Just as there are both serial and parallel aspects to processing, there exist both local and distributed aspects of storage. PDP maintains simply that the operations of the brain are *primarily* parallel and distributed in nature.

Much of what makes PDP appealing is its theoretical similarity to "neural hardware." One goal of PDP is to construct models that not only function with optimal efficiency and accuracy, but that also retain human physiological plausibility. It has prompted a rethinking of artificial intelligence and its methods, and its insights have redirected some ideas about human cognition that were based on computer models. One of these ideas regards the omission of any "central executive" or "program chip" to form and retrieve memories. To PDP, everything from learning to recall to concept formation is related to the basic, elemental units and their interactions. While this is sometimes criticized for being below the level at which a cognitive theory should operate, it is arguably a highly appropriate and logical place to start. Studying and understanding the lowest elements may lead to future discovery of "higher" functions, which would then enjoy the benefit of a strong foundation. Analogously, Rumelhart and McClelland (1986) concede that it would be difficult to understand a diamond just by studying a single carbon atom. However, it would be rather foolish to ignore what we know of this carbon atom and how it *aligns with others* when studying the diamond as a whole. Concluding simply that a diamond is "hard" is of questionable significance in understanding it.

With these ideas in mind, let us proceed to a description of the model's architecture. While various computer-simulated models differ, there are certain aspects of any PDP model that remain basically constant, outlined by Rumelhart and McClelland (1986) in their explanation of a model's general framework.

The basic element is, of course, the *processing unit* itself, equivalent theoretically to a highly-simplified model of a neuron. Each individual unit's role in the system is very limited; it is a small component of a larger system, and is basically meaningless in and of itself. The pattern resulting from the activity between many units is what defines a meaningful entity. A unit's only function is to receive input, compute an output value from it, and send that output value to other units.

Some systems use only two types of units, but many use three. Input units receive external input, either from the "world" or from other units outside the observed system. Output units direct signals from the system outside, either to other systems or directly to motoric activity. The third type of units, called

hidden units, neither influence nor are influenced by forces outside the system, but reside between the input and output units and function as modifiers of activity between the two. The earliest "connectionist-type" models, such as the ancestral "perceptron" of Rosenblatt (1958), relied only on the input and output units and were therefore somewhat limited in their capabilities. The employment of hidden units is what has allowed the more recent systems to function at a much higher level of performance.

The *state of activation* of a system reflects what a certain system is representing at a certain time, based on the activation value of its units (recall that a representation is defined by a pattern of activation across multiple units). This may be represented in vector notation by a vector **a**, with each value in that vector specifying the activation of one unit. For example, the activation of a four-unit vector **a** at time t, written as **a**(t), might be [1,0,-1,0]. This distinct pattern of activation denotes the current representation. It could represent anything—a dog, a cat, a baseball, etc.; each unit is activated to a certain value based on the stimulus input of the object, and each object in this way motivates a different "pattern of activation vector" for representing it.

Activation values may vary according to the specific model being considered. Some models specify discrete levels (which are usually binary), with "1" meaning the unit is activated and "0" meaning that it is not. A discrete model could also range from -1 to 1, as in the example above. Other models may specify the use of continuous activation values which might yield, for example, a vector such as [.6, .4, 0, -.8].

In addition to receiving inputs, units also pass output on to other units. How they affect neighboring units is a result of their own current level of activation being mapped through an output function (f) to produce an output signal $o_i(t)$ (the output of unit i at time t). Basically, a unit receives input, converts it to an output signal, and passes it along to connected units that receive it as input. Usually a threshold value is involved; that is, unless a unit is activated to a certain value, it will have no effect at all on neighboring units. This means that if a unit is activated in such a way as to compute a .6 output value, but its threshold is 1, the units connected to it receive no input from it at all.

The *pattern of connectivity* between the units represents the knowledge contained in the system. Simply, everything a system "knows" is represented by which units are interconnected, the modifiable strengths of those connections, and, as a consequence, the pattern they will produce when activated. In the simplest case, it is assumed that the input to a unit is the weighted sum of all inputs it receives from units that are connected to it. The positive or negative value of this sum determines whether the input is excitatory or inhibitory, and its absolute value denotes the strength, or weight, of the connection. More complex cases may call for different types of inputs to be summed before impinging on the designated unit. For example, all excitatory inputs may be run through one connectivity matrix to produce a value, while all inhibitory inputs are run through a separate connectivity matrix, with their final values then summed upon reaching the designated recipient unit.

The *rule of propagation* takes the "output values of the units and combines it with the connectivity matrices to produce a net input for each type of input into the unit" (Rumelhart & McClelland, 1986). This refers to the issue stated above, regarding the modification of input signals in a matrix before impinging on a unit.

The *activation rule*, expressed as the function **F**, combines all the net inputs acting on a unit with that unit's current activation level in order to produce a new level of activation for that unit. That is, the level to which a unit is activated is a function of the activation level it currently maintains and the net input of all units impinging on it.

To accommodate learning, the patterns of connectivity must be modifiable as a result of experience.

Adjustment to the interconnective weights is necessary, since it is these weights that "store" all knowledge. The most easily understood approach to accomplishing this utilizes an extended and expanded version of the Hebbian learning rule, which basically states that any time any two connected units are highly activated, the connective weight between them should be strengthened. The most commonly-used extension of this is the Widrow-Hoff, or delta rule, so named because "the amount of learning is proportional to the difference (or delta) between the actual activation achieved and the target activation provided by a teacher" (Rumelhart & McClelland, 1986). In other words, how much is learned (how much the weight is modified) is proportional to the distance between where it is and where it needs to be. The further "off-target" it is, the more it will learn. (Incidentally, the "teacher" function could be considered somewhat controversial, in that it is external in nature; however, many examples of human learning do have an external teacher that guides the formation of correct associations. This teacher, though, does not change connective weights; that is done internally in the human mind.) Symbolically, the delta rule is written $\Delta Wij = n\,(ti(t) - ai(t))\,oj(t)$ and states that the weight connection between unit i and unit j changes as a function of the proportional rate of learning, the teaching input on unit i minus that unit's current activation, and the output value of unit j.

Lastly, any PDP model must have a specified *environment* that it is to operate in. What this means is that "there is some probability that any of the possible set of input patterns is impinging on the input units" (Rumelhart & McClelland, 1986). This is significant when models are restricted as to what types of vectors they can accept as input, or what kind of stimuli they are equipped to process.

While the above aspects are in simplified form and not precisely applicable to all the various models that have been constructed (some use more complex, non-Hebbian learning rules, for example), they do provide an overview of the basic workings of a PDP system.

Builders of PDP models employ the general framework just covered, but also operate under further constraints to maintain neural plausibility in their models. One of these is the "100-step program" constraint (Feldman, 1985), regarding the aforementioned limit on the number of sequential steps a brain is capable of processing in a second. The constructed models perform comparably, time-wise, with humans; that is, if the model can produce an output in less than a second, it must use parallelism in order not to violate the "100 sequential steps per second" limit that is imposed on the human mind by the slowness of neuronal activity. The nanosecond processing capability of the computers is not allowed, for this would result in an inaccurate simulation of human performance.

A second point is brought up by PDP proponents to distance their models from the usual (or older) computer analogies to thinking. As mentioned earlier, all models consider knowledge to be stored in the connection strengths, rather than as a "state." While a pattern can be activated to a state temporarily, that state is not the knowledge. That knowledge is said to be *i*mplicit in the system, residing only in the weights. In effect, PDP wants to "replace the 'computer metaphor' as a model of mind with the 'brain metaphor' as a model of mind" (Rumelhart & McClelland, 1986).

So what are PDP models capable of? Some fairly impressive things, actually. Two models are presented in the following paragraphs in an attempt to illustrate the methods of functioning and the capabilities. Probably the simplest model to describe and understand is the pattern associator, so it will be presented first. The second model considered shows PDP's applicability to the field of artificial intelligence.

What a pattern associator does, as its name implies, is associate two different patterns of activation that are related in some manner. Rumelhart and McClelland (1986) use as an example a system that learns to associate the pattern of activation

representing the appearance of an object with that of its aroma, so that when given the visual input (e.g. [+1,-1,-1,+1]), it will produce the olfactory pattern (a similar vector).

It accomplishes this through the use of a matrix (four rows by four columns in this case), with the visual input vector placed horizontally across the top and the olfactory vector positioned vertically along the side. The goal is to have the values on the visual vector excite the corresponding value on the olfactory vector if it is a positive value, and inhibit it if it is a negative value. This is done by modifying the strengths of the connections within the matrix, so that when multiplied by the values of the visual vector, each row will sum to the desired value on the olfactory vector. For example, if the visual vector is specified as [+1,-1,-1,+1], and needs to produce a (-1) value at the top position of the olfactory vector (corresponding to the top row of the matrix), then the connective weight values would be "tuned" to [-.25,+.25,+.25,-.25]. Thus, as each value of the visual vector is multiplied by the connection strength it results in a (-.25) value; and when these four values are summed across the row they yield the appropriate (-1) value on the olfactory vector. The remaining three rows in the matrix are set up in the same way, to yield appropriate values in the remaining positions of the vertical olfactory vector. While this model may not seem simple at first glance, its simplicity becomes more evident after a short time of study and consideration (and seems even more so in light of other PDP models).

This simple model illustrates some interesting attributes of distributed representations. First, it learns through simple repetition; that is, learning is accomplished simply by repeated simultaneous presentations of the two patterns. Furthermore, the model can "teach itself" the proper set of connection weights (within the matrix), just as a result of this experience. It employs the basic Hebbian rule, extended to cover positive and negative activation values: the strength of the connection between the visual and olfactory vectors is adjusted "in proportion to the product of their simultaneous activation . . . if the product is positive, the change makes the connection more excitatory, and if the product is negative, the change makes the connection more inhibitory" (Rumelhart & McClelland, 1986). The strengths of the connections are formed gradually, in Hebbian fashion, and the information needed to determine their values is available locally from the activation of neighboring units. No "central executive" is needed. Essentially, an "empty" or "blank" pattern associator could learn to associate the two patterns simply through exposure to repeated simultaneous presentations.

An interesting point arises with regard to this pattern associator, and lends it a degree of similarity with human functioning. A perfect visual pattern is not necessary to produce an accurate olfactory pattern (though an imperfect visual pattern would result in a weaker-than-optimal olfactory pattern). For example, altering one value on the visual vector (e.g., flipping it to zero) would still result in the corresponding value on the olfactory vector being pushed in the proper direction (positive or negative). This is known as "graceful degradation," and is seen in other PDP models. It maintains, basically, that internal access to a representation is not lost completely as a result of distorted input, a characteristic that stands in stark contrast to a computer—for without a precise "address" from which to obtain desired information, a computer will produce nothing at all.

The matrix formed to associate the above two patterns can also be used to associate a separate pair of patterns, or some number of other patterns. Assume that a second pattern associator matrix is produced using the visual and olfactory patterns of another object. When this matrix is overlaid on the original one and the separate connective weights are summed at their corresponding positions, the result is a matrix that will produce the correct olfactory output given either set of visual input. Using a simple Hebbian rule as stated above, this matrix is limited to

accurately processing only inputs that are fairly distinct from one another, but matrices incorporating more elaborate learning rules are capable of handling many similar inputs. This latter type leads to the emergence of several attractive properties.

One of these is that a new visual pattern that is similar to an old one will generate a similar olfactory pattern, leading to a useful form of "spontaneous generalization." The model is also capable of extracting a "central tendency" from the repeated presentation of the same pattern with various degrees of distortion. Furthermore, the model will "recognize" and utilize regularities between different pairs of patterns, allowing the formation of interconnective strengths that produce patterns that appear to be the result of rule usage, but are really attained only through repetition of input patterns. One particular model, as an example, was fed pairs of words, the first being a root verb and the second being its past-tense form (Rumelhart & McClelland, 1986). After multiple pairs were presented, it was tested using previously unused (non-training set) words, and produced child-like errors that appeared to result from the application of rules (e.g., it converted "come" to "camed"). No "add *-ed*" rule was ever given to the system, yet its connections were formed with the tendency to apply that particular pattern.

More complex models, particularly those that utilize hidden units to modify the signals between the input and output units, are capable of substantially more impressive results. One model (Churchland, 1989) is presented here as an example, and was designed with the goal of differentiating the very similar sonar echoes of mines and rocks on the ocean floor. This is an extremely difficult task, in that the two echoes sound identical to the untrained human ear, and the echoes within each type may vary considerably.

The network was constructed with thirteen input units, each one's activation level being dependent on a certain sound frequency extracted from the echo. A layer of seven hidden units, each receiving input from all thirteen input units, then processed the incoming signals and sent modified output signals to the two output units. Only two output units were needed, for the ideal output values of [1,0] for mines and [0,1] for rocks.

The model learned by being given multiple examples of mine and rock echoes (vectorized by sound frequency) one at a time, and having its output evaluated after each. Following each presentation, its actual output was compared with its target output, and the connective weights within the system that were deemed most likely to be causing the error were adjusted according to Rumelhart, Williams, and Hinton's algorithmic "generalized delta rule" (Rumelhart & McClelland, 1986). After several thousand presentations and weight modifications, the system was surprisingly accurate in its ability to distinguish the two types of echoes in its training set (approximately 90% correct). It had "tuned" its connections to detect whatever combination or pattern of features was unique to each type of echo. The "knowledge" was in its connections weights and was obtained through a learning algorithm that allowed those weights to be acquired.

What was going on inside the model to allow the connection weights alone to distinguish a rock echo from a mine echo? One must stretch, intellectually, and grasp the notion of a seven-dimensional "hyper" figure within, an abstract space with one dimension supplied by each of the hidden units. Each echo vector that was fed into the system, and consequently to the hidden units, fell into one point in this "hidden unit activation vector" space. The system's objective was to divide this space into two subspaces, one to represent rocks and one to represent mines. The output units' only job was then to determine which subspace a given echo fell within.

Yet this system's capabilities go impressively further. Echo vectors that occupied the center of each subspace were the prototypes and would produce output values near the goals of either [1,0] or [0,1]. Those vectors that fell near the boundary separating the two subspaces

would retain the proper relationship between the two values, but would be less precise--for example, [.6, .4]. This illustrates the system's "graded responses," and is evidence of its sensitivity to similarities across the dimensions. The system is impressive in and of itself, but becomes more so when its characteristics are applied to something like human speech perception, where we accurately process a highly variant set of phonemic input into "subspaces" of correct categorization. Interestingly, this human capability was simulated to some degree in the PDP model of NETtalk (Rosenberg & Sejnowski, 1987).

Moving into the construction of memory models, PDP focuses once again on the micro-elements, exploring and theorizing on the roles of neurons and their connections. By the theorists' own admission, their memory model does not at this point attempt to elucidate the detailed processes involved in the retrieval and use of memory to guide behavior. What they are attempting to do is construct a physiologically plausible model that conforms to empirical data that strongly imply the storage of both general and specific information, in the form of abstracted prototypes and specific exemplars. This is done within the general framework of individual units, activation values, and weighted connections.

In the memory model, the units are structured into larger entities called modules, each of which consists of a large number of interconnected units. The modules themselves are also interconnected, and each may receive input either from the other modules or from external stimuli.

A mental state is defined as "a pattern of activation over the units in some subset of the modules" (Rumelhart & McClelland, 1986). Basically, the pattern currently activated represents what resides in conscious thought at that moment. Each memory, or pattern of activation, leaves a trace, or slight change in the interconnective weights. Being of a distributed nature, all memory traces in the system leave their influence in a common set of weights. The process of retrieval is simply the calling-up of a prior mental state with the aid of an externally-supplied cue, which would itself be some part or "fragment" of the original state. This external cue could possibly come either from the "world" or from an internal search process, although a specific mechanism for the latter has not yet been fully developed for incorporation into the model.

The units in the model can take on any value between -1 and 1. A zero value is considered neutral, and weights will tend to decay toward it with time. In an attempt to maintain consistency with human functioning, this decay is assumed to be rapid at first and then gradually slowing. This translates roughly into the "freshness" of a recent human memory, and how it becomes less easily accessible with the passage of time.

Each memory trace that passes through the module causes a slight change in the complete set of weights. The delta rule is utilized for establishing the correct direction and magnitude of each weight change, and its function is to enhance the "storage of connection information." That is, given a partial pattern as a retrieval cue, the connections between the units are weighted in such a way as to reinstate the complete pattern by providing appropriate excitatory/inhibitory input to the connected units. The goal is to have internal activation match external activation. For example, if a certain unit is excited by some input pattern, the connections leading to it from other units will tend to excite it as well.

The specific model explained in the following paragraphs (Rumelhart & McClelland, 1986) was designed to illustrate what occurs in the storage of memories, and how memories naturally categorize themselves and form a prototype for that category. It shows how multiple concepts can be held in the same set of weights, and the ability of specific examples to exist in the weights in addition to the prototype.

The model was constructed using twenty-four units—sixteen for representing the visual patterns of

activation for different "dogs," and the remaining eight for representing various different names, or types, of dogs. A prototype "dog" vector was constructed, being just a randomly-chosen string of -1 and +1 values. This prototype was never presented to the model, but fifty random distortions of it were, each obtained from the random flipping of different values on the vector. After each presentation of a distortion, the delta rule modified the weights appropriately, and these increments to the weights were then allowed to decay down to about five percent of their initial effect before the next presentation.

What results is a set of weights that changes slightly with the presentation of each new distortion, but after fifty presentations of distortions it forms a matrix remarkably close to the "prototype dog" pattern. Most deviations from the prototype are caused by the most recent distortions presented to it. This is due to the lesser amount of time to decay and the fact that no further presentations have been presented to "blend" their influence. This is consistent with observed characteristics in human memory. The model will more easily "recognize" the prototype than any specific examples (distortions), but it cannot apply a name to the *prototype* pattern, since each distortion was given with a different name vector (corresponding to the different types of dogs). If, however, all the examples given had been named simply "dog", then the model would produce the prototype visual pattern in response to the prompt of that name (and vice versa--if given the prototype visual pattern, it would respond with the name pattern for "dog").

Yet the model's abilities go beyond this. The model was next shown to be capable of storing three different concepts in the same set of twenty-four units. Two of the patterns were similar (nonorthogonal), representing "dog" and "cat," while the third pattern was orthogonal and represented "bagels."

When given distortions of the three different prototypes (under the same procedure as above), and given the appropriate category name simultaneously, the model assigned each pattern to its proper concept. That is, after a sufficient number of training trials, the model was able to produce the appropriate prototype pattern when given a category name. As with humans, making the distinction between dogs and cats took longer than distinguishing those two from bagels, but the delta rule ultimately settled on a set of weights that reliably produced the desired results. The model accomplished this by the fiftieth trial; all three concepts and their prototypes were held within the same set of weights. Furthermore, the model was proven able to form these three visual-pattern categories without the aid of being given names; in this case, the eight-valued name vectors presented with the visual patterns were all flipped to zero. Still the model achieved proper categorization of the three different patterns, with no mechanism supplied for doing so.

The model can also store the patterns of specific exemplars in addition to the prototype, as is obviously necessary in human memory. In this situation, the model was again given fifty training trials. One particular distortion of the prototype was given the name "Rover" as an eight-unit name pattern vector, another was given "Fido," and the remaining distortions were simply called "dog." After training, the model could produce the appropriate visual pattern given the name of either exemplar (Rover or Fido); if given the name "dog," it would produce the prototype visual pattern. Its ability to do this lies in the weight connections. The name pattern for "Rover" is connected with the visual pattern in such a way as to reinstate the pattern. Incidentally, there was only one element that differed between the "Rover" visual pattern and the prototype pattern. Because of this, the unit representing that element had an extremely strong connection to the pattern representing the "Rover" name.

What we see regarding the "Rover" illustration is "content addressability," which means that a memory is accessible given a partial element of its pattern as a prompt. This

prompt, or retrieval cue, is not used as an aid in "finding" the whole memory in permanent storage and bringing up a copy of it to working memory; rather, it is a *part* of the memory itself, and its activation will complete the activation of the entire pattern via interconnections between units.

The basic ideas underlying this model can be applied to the explanation of other aspects of memory as well. Models have been constructed that can accurately duplicate the effects of semantic priming seen in human behavior (Rumelhart & McClelland, 1986). Semantic priming—the ability to recall a familiar item more easily and quickly if it has been recently experienced—has fallen into some difficulty lately in its interpretation. Traditionally, the priming effect seen in word recognition was explained by assuming the presence of a "word detector" or "logogen" in the mind, with a threshold level for activation. Each time the word detector was activated, the threshold would be lowered (temporarily—eventually the activation effect decays). But Rumelhart and McClelland (1986) note that empirical testing (Jacoby, 1983) has shown that changes in context (e.g., different voices or media employed in the priming and test conditions) result in weaker priming effects, which would not be predicted if dealing with a single word detector. Presumably, any manner in which said word detector is accessed should serve equally well.

The PDP model used to account for this strays little in concept from the above theory, but attempts to describe priming effects physiologically and to account for the problem of changing contexts. Imagine a stimulus detector spread over a set of weights, such as the structure of the "dog" module presented above. Each activation of that particular stimulus (assume for convenience that the stimulus is a word) contributes to a composite memory trace, much like all the different presentations of "dogs" contributed to an averaged prototype; but "the characteristics of particular experiences tend nevertheless to be preserved, at least until they are overridden by canceling characteristics of other traces" (Rumelhart & McClelland, 1986).

The ability to more easily perceive a recently-presented stimulus is a result of the composite memory trace being recently active, allowing a new presentation of the stimulus to settle into a stable pattern of activation across the units more quickly. The effect of changing contexts is explainable by the supposition that recent memory traces have not decayed as much, leaving a relatively strong influence on the pattern across the units. If a stimulus is presented in a testing condition that is identical to the priming situation, these strong traces enable a match to be made promptly. If the stimulus is presented under different testing conditions, the system does not have the benefit of an exact match and must access the somewhat "distorted" averaged pattern that resides in the composite memory trace as a whole. What this type of structure in a model does is continually update a "summary representation," with the most recent exemplars influencing the pattern slightly in their direction. With time, or with subsequently presented exemplars, the characteristics of the exemplars blend into a composite memory trace. The more similar each exemplar is to the summary representation, the more they will blend, and lose their individual characteristics relating to the context in which they were experienced. Exemplars that vary widely from the central tendency will tend to be held as memory traces in the context in which they were experienced.

The above point can be used to explain the gradual conversion of episodic memory traces to semantic memories (if one subscribes to a distinction between the two). Assume a "proposition" is experienced at several different times and in several different contexts. If each experience is generally consistent with the others, a central tendency or summary representation will emerge from the similarities between them, while their unique contexts will gradually "wash out" as a consequence of their evident

irrelevance. What is left is a summary representation of a proposition that is considered semantic in nature, but that was formed naturally by episodic experiences.

Furthermore, this line of thinking can be employed to explain the results of the well-known Loftus experiments (e.g., Loftus & Palmer, 1974), in which subjects' memories were apparently "changed" as a result of the wording that was used when asking for recall. PDP would contend that one memory was not "overwritten" by the other, but that the wording of the recall question formed another memory "trace" that blended with the first one. Since, according to PDP, memories do not exist as a whole but as a set of weights that are modified with each new trace, the observed results in the Loftus experiments reflect a modified composite memory. Could the "original" memory be drawn out, as has been suggested, through the use of hypnosis or other techniques? Probably, if the original trace has not decayed sufficiently into the blend, and can be reinstated with strong enough contextual fragments of its original pattern as cues.

What PDP is attempting to show with its memory models is that simple units, in conjunction with the delta rule for modifying weights between them with information that is locally available, can account for observed characteristics of memory without any type of executive overseer. They will automatically and naturally draw out a central tendency from input without any mechanisms for generalization and without any rule applied by a "program chip." PDP does not assert that rules do not exist in any area of cognition, but merely intends to show that some processes that appear to use rules or a "program" are not necessarily doing so.

How does PDP's conception of memory compare with other theories? It obviously differs considerably from the original stage models based on computers, such as Atkinson and Shiffrin's (1968). While this type of model does not enjoy the acceptance today that it did upon its introduction nearly three decades ago, its abstract framework and its distinction between "working" and "permanent" memory are still readily referred to and utilized in explaining cognitive processes.

PDP does not subscribe to this division of memory, its primary objection being to envisioning working memory as a "central processor" or an executive, capable of searching a "file cabinet" of separate and complete memories in permanent storage. As stated earlier, PDP does not view memories as "whole" entities, but simply weights in the connections between units. For PDP, the "file cabinets" would have to be somehow interconnected with drawers that slid into and through others. The documents in "one drawer" may be the "A's" when accessed once, and the "P's" when accessed the next time. Moreover, new documents added to a drawer would tend to blend with the other ones, changing the characteristics slightly of all the documents previously in that drawer, which themselves had already influenced each other. Stretching this ill-fated analogy further, the file cabinets would even be "empty" until a the need for a document was prompted, at which point some number of them would suddenly produce it with no search necessary. PDP, despite its contentions otherwise, appears to be fairly passive in many areas and does not at this point allow an "active processor" to guide a memory search.

Before the more recent (and more empirically plausible) models are examined a point should be made. The authors of PDP in its original form maintain that their theory is not necessarily to be viewed as an alternative to other models. That is, it does not proclaim the other models to be necessarily incorrect in their assumptions of the structures and workings of memory. Rather, PDP is interested in providing a neuronally-plausible explanation for what may be occurring within these models. PDP is in essence attempting to dissect these abstract, symbolic processes and structures and provide a physical basis for them. When and if these physical processes are sufficiently exposed to the satisfaction of a majority, they may lead to

altered versions of these existing models (or completely new ones), but will not necessarily require them. In fact, they may end up supporting and strengthening these previous theories instead. With this in mind, a few other models can now be examined, noting what PDP has to say about their assertions.

Craik and Lockhart's (1972) levels of processing approach states that the more deeply material is processed, the more durable its resultant memory trace will be and the more easily it will be retrieved. "Depth" basically correlates with semantic processing, as compared to "shallow" processing of a stimulus' physical characteristics. It shares the assumption with PDP that memory is not composed of separate storages.

Experimental support for the level of processing as a determinant of recall ability (Craik & Watkins, 1973; Parkin, 1984), wherein subjects better recalled material that they had semantically processed, could be explained by the contentions of PDP. The point could be made that semantic processing—as opposed to the judging of the physical characteristics of a stimulus word—adds more dimensions (employs more units) and results in a more richly interconnected pattern. In effect, semantically processing a stimulus word would usually involve relating it to its context or associating it with related things. Each would form connections from the stimulus to the related items, meaning that there would be a much broader base from which to draw a partial pattern, which would then serve as an impetus for reinstating the whole. The larger a prompted fragment is, the better able it is to excite and inhibit its interconnected units and achieve pattern completion. Interestingly, this explanation can account for the data that contradict levels of processing as well. One experiment that produced such data found subjects recalling the stimulus word "dime" equally well when given the physical-feature cue "ime" as when given the semantic cue "an American coin" (Nelson & McEvoy, 1979). Again, it is a matter of pattern completion. The subject has seen the word "dime" and it has registered as a visual pattern of activation across interconnected units. When given "ime" as a cue, seventy-five percent of its word detector units are activated and can each send appropriate activation to the unit responsible for representing the "d" in the first position. Thus the pattern is completed easily by virtue of internal connections formed during exposure to the stimulus.

The teachable language comprehender (TLC) of Collins and Quillian, possibly the first well-known model to employ a hierarchical structure of memory items, suffered from test results that were inconsistent with its predictions. One of its major problems stemmed from subjects' tendencies to violate the hierarchy, as in more quickly confirming a dog's status as an animal than its status as a mammal (Rips, Shoben, & Smith, 1973). PDP, of course, denies any hierarchy in the structure of memory. Its evaluation of this effect would center on connection strengths and would be explained by more heavily-weighted connections between "dog-animal" than between "dog-mammal." This is due to the principle outlined in the basic Hebbian rule, that connective weights are strengthened with more associations or simultaneous activations. Most people hear "dog" associated with "animal" much more frequently than with "mammal"; it would be rare to hear someone ask if there were any mammals in your backyard. Judging from the literature, frequency of association was generally suspected to be the culprit in this, and the Hebb rule provides a comfortable degree of support for this conclusion.

Another problem the TLC model had regarded "better examples" of categories (Smith, Shoben, & Rips, 1974). For example, subjects might verify a peach more quickly than a watermelon as a fruit. It is possible to account for this finding with the assertions of PDP as well. When the word "fruit" is heard, assume it triggers the prototype pattern of activation for that concept (imagine it as a vector of some length with various positive and negative activation values). The activated

pattern for "peach" closely resembles the pattern of this prototype, because its features closely resemble the "averaged" pattern of all fruits. A watermelon's pattern, however, does not as closely resemble the averaged prototype, and would thus take longer to verify. The existence of a prototype was never in doubt in explaining the effect of the "better example," but PDP provides a attractive account of the underlying processes involved.

Collins' next attempt, the spreading activation model (Collins & Loftus, 1975), dropped the hierarchical structure and replaced it with nodes whose associative strengths with other nodes were reflected in the lengths of the links between them. This type of model begins to approach the thinking of PDP. PDP would use connective weights to signify associability rather than semantic distance along a link and would further attempt to expose the structure within the nodes (as a pattern of activation) rather than accepting them as abstract "wholes." Spreading activation assumed serial processes as opposed to PDP's parallel, but the remainder of its assumptions (regarding the strength of activation, and the more or less involuntary activation of closely associated nodes) could be adapted and modified to fit within the PDP approach without too much difficulty. A weakness of the spreading activation model lay in the "prestored knowledge" supposed to exist in the structure which readily supplied "no" answers via "is not a" links. PDP would never assume a priori knowledge, and would thus avoid such a weakness (although negative answers pose an equally formidable challenge for PDP models).

Anderson's Adaptive Control of Thought (ACT) theory (1976) is possibly the most sophisticated in the realm of semantic networks. It also, with a few exceptions, approaches the principles embodied in PDP. The storage of propositions is likely an apt way of representing knowledge, though Anderson fails to speculate on the physiological aspects of how these propositions are constructed. The propositions could be likened to memory modules in a PDP system, and their interconnective links representing the strengths of the associations between them could easily be explained with connective weights.

Yet ACT strays from PDP as well. As with previous models, Anderson speaks of adding new propositions that form links to existing ones through an ever-growing array of extensions and nodes. It is somewhat analogous to building a structure on a table with wooden blocks and dowels. Each new proposition is represented by a wooden block, which must be properly connected with dowels to all other blocks it is associated with. The task of making all applicable connections with every new block soon becomes formidable and is restrained by the physical nature of the structure. It would be much simpler to have, say, a small box on that table with an electric grid. Each new proposition could be sent through the same grid as a pattern of impulses, and the associations and generalizations formed would be implicit. Such is the orderly structure that PDP imposes on the storage of knowledge.

Another deviation from PDP in the ACT model regards its use of rule-based "production systems." PDP, while not ruling out the use of rules in cognition, prefers to search for explanations that do not necessitate them. A PDP model would never employ the use of production rules as a starting point in constructing a model, as this would tend to imply executive functioning.

An interesting aspect of the relationship between episodic and semantic memory is brought up by Best (1992) in his discussion of the ACT theory. The discussion centers on the "type" node for cats (general, semantic memory) and the "token" nodes (specific, episodic memory) that branch from it. He states that in his own token nodes under "cat" there would be the representation "my cat recently scratched the sofa". He proceeds to theorize that with a few more observed examples of sofa-scratching by cats this episodic fact could be generalized and elevated to the level of semantic

knowledge in the type nodes (things that are true of all cats).

PDP would take a different viewpoint on this, specifically in the need for "a few more examples of sofa-scratching by cats". In PDP's view, one single episode of sofa-scratching would enter the weights relating to cats and would tend to generalize, at least to some degree. The result would be that a person in this situation would judge sofa-scratching to be a trait of cats in general, albeit with a degree of probability low enough to spur the testing of this hypothesis. This is because, in effect, PDP's equivalent of "type" and "token" nodes lie within the same set of weights. In other words, each "cat" exemplar (and the weights that represent it) lies within a system of weights representing the overall concept of a cat. Under this supposition, if Best never again encountered any other cat scratching a sofa, the weights in the module representing "his cat" would be adjusted to reflect that sofa-scratching was unique to that one exemplar. More likely, he would encounter a number of subsequent examples pertaining to the scratching of a sofa by a cat, each of which would augment the weight representing that in the overall cat pattern.

While PDP offers some enticing explanations of some aspects of cognitive functioning, it is obviously not universally accepted. As stated before, it is a considerable departure from conventional thinking, and therefore must answer to a wide range of questions that inevitably arise from the field. Presented in the following paragraphs is a sampling of some such objections and concerns.

One of the objections deals with the hidden units. These aforementioned units reside in a layer between the input and output units and transform the signals between them in such a way as to greatly increase the processing power of a PDP model. While they may possibly be physiologically equivalent to interneurons in the mammalian nervous system, which are neither sensory nor motor (Hanson & Burr, 1990), some aspects of their incorporation into models have been questioned. Among these are the concerns regarding how to determine the number of hidden units that are to be employed in a given model, and whether this chosen number is based on any neurological data. Furthermore, there is some question about how the connective weights are assigned to the hidden units, with some suggesting the possibility of post-hoc fit with the data. In summary, one feeling is that "the number of hidden units, their connectivity with other units, [and] the weights...should be justified in terms of explicit principles" (Haberlandt, 1990).

PDP's response to this is based on the fact that it is still in its early stages and that, due to considerable self-imposed constraints on its structure, "there are no obvious principles that will allow the generic design of connectionist (PDP) models at this point" (Hanson & Burr, 1990). Modeling systems that will function acceptably under the given constraints is the primary focus at this time; perhaps specific rules governing the many aspects of hidden units will emerge with continued research.

Another objection concerns the issue often discussed in PDP of learning rules for behavior versus learning behavior through repetition that only appears to use rules. It has been argued that taking away rule formation leaves a substantial gap in human reasoning. In other words, it is our ability to form and utilize rules, rather than relying on the "percentages" predicted by repetition, that places humans above the lower animals on the reasoning ladder. Knowing the rule for obtaining B from A is vastly superior to just knowing that B results from A (Hendler, 1990). Levelt (1990) adds that a human, if told by the phone company to "add 2 to the end of every phone number starting today," could apply that rule easily without having to be retrained on every phone number in his memory, as a PDP system would theoretically require.

Levelt goes on to criticize the notion that learning in PDP networks is "natural," i.e., closely resembling human learning. He argues that models constructed in PDP are limited and specialized, thereby not allowing them to

be compared to human cognition from a psychological standpoint. One of his chief criticisms concerns the overriding of pre-existing knowledge when acquiring new knowledge. A child learning arithmetic, he maintains, can be taught addition first and multiplication later, without having to be retrained on the former. The only way to achieve both skills in a PDP model, he says, is to "train up" the two operations simultaneously, which does not correlate well with human behavior.

PDP counters this line of thinking on two related fronts, both dealing with the limited abilities of constructed models at the current time. First, regarding rule usage, PDP maintains that a system "adopts representations to perform one task (as far as it knows) and if 'enough' constraints are present the network can apply the representation to new tasks" (Hanson & Burr, 1990). This is a question of "scaling," really; that is, the difficulties in mapping to new tasks are based on expanding the domain of the model, and not on any inherent weakness in the theory. Furthering this thought in a natural extension to the second point, "one must realize that when there is a dissociation between technology and theory, it is easy to make bad models" (Hanson & Burr, 1990). Simply stated, technological limitations need not damn the underlying theory; transforming a complex theory to hardware is a trial-and-error endeavor at times, and its current lack of success should not be interpreted as a failure of the theory itself.

PDP has been criticized by some, especially those in the field of linguistics, for conveying a system too close to empiricism, or "blank slate" views on mental development. Indeed, Hanson & Burr (1990) downplay Chomsky's Language Acquisition Device, stating that any tendencies a human is likely to be born with would more probably deal with extreme generalities, such as the three-dimensionality of the world, alternating cycles of light and darkness, etc., and not to "specific activities (such as chess playing, tennis, or speaking English, Shepard, 1987)." However, opponents of this argue in favor of something at least akin to LAD, employing as evidence the rapid rate with which we learn language as compared to other, less complex tasks. They argue that there must be some sort of naturally-constrained, "prewired" learning process to permit this (Jordan, 1990). Such a mechanism does not fit well within the framework and ideology of PDP.

Others question the value of PDP's contribution to the field of psychology. While hidden units may have a neurological equivalent, the functioning they allow may be inconsistent with observed functioning in humans. Lamberts and d'Ydewalle (1990), for example, argue that if experimental evidence implies the use of three processing stages for a person to compute a certain mapping, and if a PDP model with hidden units is constructed to do it in only two stages, then that model is simply incorrect psychologically. Lamberts and d'Ydewalle further contend that simply because a model has "neural plausibility" does not necessarily make it psychologically relevant. That is, that a model can accurately explain what the brain *could* be doing within the constraints of physiological knowledge cannot be taken without question to mean that this is what it actually *is* doing (an application of the fact that a theory's successful explanation of something does not guarantee the correctness of that theory).

Again, PDP is forced to stand behind its early stage of development in answer to this, implying that it is perhaps not yet "time to map these models into specific theories of cognition" (Hanson & Burr, 1990). Groundwork is being laid and its success is encouraging, but PDP admits that it is not yet a perfect match to biology. However, given the restrictions it places on itself in regard to neural architecture, it seems the future application of the beliefs to specific cognitive theories is promising.

Some in the field of psychology argue that all representations in a PDP network are symbolic, basically being the transformation of external stimuli to excitatory or inhibitory impulses between units, and that this type of representation

is insufficient for interacting with the world. Icons, they say, are necessary to preserve the internal structure of the stimuli that form them, to preserve "similarity relations," and to allow for generalization. Symbolic representations are only arbitrarily related to the external stimuli that form them, and can therefore "neither preserve similarity relations nor support generalization" (Phillips, Hancock, and Smith, 1990).

This is countered somewhat by the working example of the model discussed earlier (that dealt with the recognition of "dogs"), wherein (what is here referred to as) a symbolic system was indeed successful in "preserving similarity relations," extracting a prototype, etc. Churchland (1989) states that the brain is "a purely physical system...short of appealing to magic, or simply refusing to confront the problem at all, we must assume that some configuration of purely physical elements is capable of grasping and manipulating features, and by means of purely physical principles." Neurons compose the brain and are relatively simple entities, their duties apparently being merely the sending and receiving of impulses.

Other objections focus on the complexity of PDP, and the necessity of searching for an irreducible element. Suppes (1990) contends that the barriers to understanding all the intricate details of a highly complex process (such as human cognition) are functionally insurmountable, and the issue is better studied at a higher, more general level. The argument has also arisen regarding our lack of knowledge about what is happening *within* the model itself. Pavel (1990) warns that "we must be wary of modeling one complex system that we do not understand (e.g., Homo sapiens) by another (e.g., [PDP] networks)."

There are other objections that could be discussed, among them the assertion that PDP is passive and therefore not cognitive, that there is not enough known about neurology to build models that attempt to simulate the brain, etc., but detailing all of them goes beyond the scope of this paper. It should be noted, though, that proponents of PDP have plausible responses to a large number of these objections, some of which have been touched on in other sections of this paper. Until PDP is further developed, there will be remain some controversial points, particularly on the philosophical front.

PDP is enduring well, given the boldness of its assertions. Like a new theory in any discipline that is a considerable departure from the currently accepted one, it has to constantly defend itself; yet it has not seemed thus far to have suffered quite so much of a vehement and universal denial as other frontier theories have. Perhaps this reflects a new cautiousness in the scientists of today in their propensity for rejecting radical new ideas (some of whom in psychology may still be reminded of the complete paradigm shift away from behaviorism), or perhaps it is a result of its "naturalness." By this I mean that there probably exist a significant number of people in the field of cognition that have felt some degree of discomfort in accepting and referring to cognition as an abstract process. The brain is essentially the last organ to be understood physically, and any progress on that front tends to lend credibility to the field and creates a firm foundation on which to build, not one that may be swept away by the next theory that arises. In other words, many people are ripe for the study of cognition from a physiologically-supportable standpoint, as opposed to a metaphorical one.

Furthermore, PDP may well allow some previous theories back into at least partial acceptability. Much of how it accounts for various cognitive operations dances dangerously close to behaviorism (which it has, in fact, been accused of), and its ideas about "settling" or "relaxing" to solutions regarding schemata would probably induce a smile from the ostracized gestalt theorist. Perhaps PDP has the ability to marry the past, present and future into one collective viewpoint, extracting the accepted strengths of past theories and incorporating them into a new weave.

References

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, NJ: Erlbaum.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In W. K. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 1, pp 89-195). New York: Academic Press.

Best, J. B. (1992). *Cognitive psychology* (3rd ed.). St. Paul, MN: West.

Churchland, P. (1989). *A neurocomputational perspective.* Cambridge, MA: MIT Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review, 82,* 407-428.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

Craik, F. I. M., & Watkins, M. J. (1973). The role of rehearsal in short term memory. *Journal of Verbal Learning and Verbal Behavior, 12,* 599-607.

Feldman, J. A. (1985). Connectionist models and their application: Introduction. *Cognitive Science, 9,* 1-2.

Haberlandt, K. (1990). Expose hidden assumptions in network theory. *Behavioral and Brain Sciences, 13,* 495-96.

Hanson, S. J., and Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences, 13,* 471-489.

Hendler, J. (1990). But what is the substance of connectionist representation? *Behavioral and Brain Sciences, 13,* 496-97.

Jacoby, L. L. (1983). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 9,* 21-38.

Jordan, M. I. (1990). A non-empiricist perspective on learning in layered networks. *Behavioral and Brain Sciences, 13,* 497-98.

Lamberts, K. and d'Ydewalle, G. (1990). What can psychologists learn from hidden-unit nets? *Behavioral and Brain Sciences, 13,* 499-500.

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

Levelt, W. J. M. (1990). On learnability, empirical foundations, and naturalness. *Behavioral and Brain Sciences, 13,* 501.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13,* 585-589.

Nelson, D. L., & McEvoy, C. L. (1979). Encoding context and set size. *Journal of Experimental Psychology. Human Learning and Memory, 5,* 292-314.

Parkin, A. J. (1984). Levels of processing, context, and facilitation of pronunciation. *Acta Psychologia, 55,* 19-29.

Pavel, M. (1990). Learning from learned networks. *Behavioral and Brain Sciences, 13,* 503-504.

Phillips, W. A., Hancock, P. J. B., & Smith, L. S. (1990). Realistic neural nets need to learn iconic representations. *Behavioral and Brain Sciences, 13,* 505.

Rips, L. J., Shoben, E. B., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12,* 1-20.

Rosenberg, C. R., & Sejnowski, T. J. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1,* 145-168.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for

information storage and organization in the brain. *Psychological Review, 65*, 386-408.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing*, Volumes I and II. Cambridge, MA: MIT Press.

Shepard, R. N. (1987). Toward a universal law of generalization for cognitive science. *Science, 237*, 1317-1323.

Smith, E. E., Shoben, E. J., & Rips, L. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81*, 214-241.

Suppes, P. (1990). Problems of extension, representation, and computational irreducibility. *Behavioral and Brain Sciences, 13*, 507-508.