From Embodied Cognitive Science To Synthetic Psychology

Michael R.W. Dawson
Biological Computation Project, University of Alberta
mdawson@ualberta.ca

Abstract

One new tradition that has emerged from early research on autonomous robots is embodied cognitive science. This paper describes the relationship between embodied cognitive science and a related tradition, synthetic psychology. It is argued that while both are synthetic, embodied cognitive science is anti-representational while synthetic psychology still appeals to representations. It is further argued that modern connectionism offers a medium for conducting synthetic psychology, provided that researchers analyze the internal representations that their networks develop. Some case studies that illustrate this approach are presented in brief.

***Keywords**: embodied cognitive science, synthetic psychology, connectionism*

1. EMBODIED COGNITIVE SCIENCE

Cognitive science is an intensely interdisciplinary study of cognition, perception, and action. It is based on the assumption that cognition is information processing [1], where information processing is generally construed as the rule-governed manipulation of data structures that are stored in a memory. As a result of this assumption, a basic aim of cognitive science is identifying the functional architecture of cognition – the primitive set of rules and representations that mediate thought [2].

Of course, not all researchers are comfortable with adopting this research program, because they have fundamental disagreements with this foundational assumption. For example, starting in the early 1980s many connectionists argued against the need to define information processing in terms that require explicit rules and representations [3, 4]. They pushed instead for a form of information processing that is more analog and more biologically plausible.

Another tradition of research has arisen in reaction to classical cognitive science in recent years, and has been associated with a variety of labels. These include behaviour-based robotics [5], new artificial intelligence, based-based artificial intelligence, and embodied cognitive science [6]. The embodied cognitive science movement is gaining popularity, and is challenging the traditional symbol-based conception of artificial intelligence and cognitive science along many of the same lines that were adopted by connectionist researchers in the early 1980s. Embodied cognitive science is a reaction against the traditional view that human beings as information processing systems "receive input from the environment (perception), process that information (thinking), and act upon the decision reached (behaviour). This corresponds to the so-called sense-think-act cycle" [6]. The *sense-think-act cycle*, which is a fundamental characteristic of conventional cognitive science, is an assumption that the embodied approach considers to be fatally flawed.

Embodied cognitive science argues that theories of intelligence should exhibit two basic characteristics. First, they should be embodied, meaning that the theory should take the form of a working computer simulation or robot. Second, they should be situated, meaning that the simulation or robot should have the capability of sensing its environment.

Why are these two properties fundamental? The answer to this question emerges from considering the answer to a second: From where does the complexity of behaviour arise? Simon [7] imagined an ant walking along a beach, and that its trajectory along the beach was traced. Accounting for the behaviour of the ant would be equivalent to explaining how the many twists and turns of this function arose. One might be tempted to attribute the properties of this function to fairly complicated internal navigational processes. However, Simon

Proceedings of the First IEEE International
Conference on Cognitive Informatics (ICCI'02)
0-7802-xxxx-x/02/$10 © 2002 IEEE

pointed out that this would likely lead to an incorrect theory. "Viewed as a geometric figure, the ant's path is irregular, complex, hard to describe. But its complexity is really a complexity in the surface of the beach, not a complexity in the ant" (p. 51). In other words, fairly simple dispositions of the ant – following the scent of a pheromone trail, turning in a particular direction when an obstacle is encountered – could lead to a very complicated trajectory, if the environment being navigated through was complicated enough.

Embodied cognitive scientists create embodied, situated agents in order to take advantage of exactly this type of emergence. One of the aims of embodied cognitive science is to replace the sense-think-act cycle with mechanisms of sensory-motor coordination [6] that might be construed as forming a *sense-act cycle*. The purpose of this change is to reduce, as much as possible, thinking -- the use of internal representations to mediate intelligence. What makes this a plausible move to consider is the possibility that if one situates an autonomous agent in the physical world in such a way that the agent can sense the world, then no internal representation of the world is necessary. "The realization was that the so-called central systems of intelligence – or core AI as it has been referred to more recently – was perhaps an unnecessary illusion, and that all the power of intelligence arose from the coupling of perception and actuation systems" [5].

1.1 Historical Examples Of Emergence

Embodied cognitive science is an attractive approach, because it can call on a long history of success stories in which extremely interesting behaviours emerged from fairly simple devices.

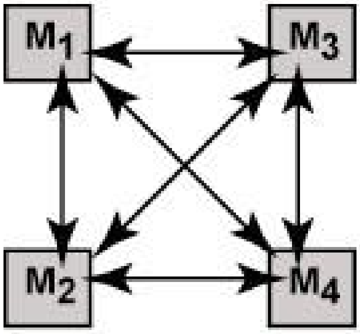

Figure 1. Feedback relationships between four different machines.

1.1.1 The Homeostat. One important historical example of emergence comes from the study of feedback interactions between generic machines by Ashby [8]. For Ashby, a machine was simply a device which, when given a particular input, generates a corresponding output. Of particular interest to Ashby was a system of four different machines coupled together with feedback, as is shown in Figure 1. Ashby [9] makes the following observation about a system of this complexity: "When there are only two parts joined so that each affects the other, the properties of the feedback give important and useful information about the properties of the whole. But when the parts rise to even as few as four, if every one affects the other three, then twenty circuits can be traced through them; and knowing the properties of all the twenty circuits does *not* give complete information about the system."

How, then, can the behaviour of such a system be studied? Ashby [8] dealt with this question by constructing a device, called the homeostat, that allowed him to observe the behaviour of this complicated set of feedback relationships.

The homeostat was a system of four identical component machines. The input to each machine was an electrical current, and the output of each machine was also an electrical current. The purpose of each machine was to transform the input current into the output current. This was accomplished by using the input current to change the position of a pivoted magnet mounted on the top of the component. In essence, each machine output an electrical current that was approximately proportional to its needle's deviation from its central position. All things being equal, a large current that was input to the component would cause a large deflection of the magnet (and needle), which in turn would result in a proportionately large current being output.

The four units were coupled together to create a system of the type that was drawn in Figure 1. Specifically, the electrical current that was input to one unit was the sum of the electrical currents that was output by each of the other three units, after each of these three currents was passed through a potentiometer. The purpose of the potentiometer was to determine what fraction of an input current would be passed on to deflect the magnet, and thus each potentiometer was analogous to a connection weight in a PDP network. The result of this interconnectedness was a dynamic system that was subject to a great deal of feedback. "As soon as the system is switched on, the magnets are moved by the currents from the other units, but these movements change the currents, which modify the movements, and so on" [8].

In order to dictate the influence of one unit upon another in the homeostat, one could set the resistance value of each potentiometer by hand. However, Ashby [8] used a different approach to allow the homeostat to automatically manipulate its potentiometers. Each unit was equipped with 25-valued uniselector or stepping switch. Each value that was entered in the uniselector was a potentiometer setting that was assigned randomly. A unit's uniselector was driven by the unit's output via the deflected needle. If the output current was below a pre-determined threshold level, the uniselector did not activate, and the potentiometer value was unchanged. However, if the output current exceeded the threshold, the uniselector activated, and advanced to change the potentiometer's setting to the next stored random resistance. With four units, and a 25-valued uniselector in each, there were 390,625 different combinations of potentiometer settings that could be explored by the device.

In general, then, the homeostat was a device that monitored its own internal stability (i.e., the amount of current being generated by each of its four component devices). If subjected to external forces, such as an experimenter moving one of its four needles by hand, then this internal stability was disrupted and the homeostat was moved into a higher energy, less stable state. When this happened, the homeostat would modify the internal connections between its component units by advancing one or more of its uniselectors to modify its potentiometer settings. The modified potentiometer settings enabled the homeostat to return to a low energy, stable state. The homeostat was "like a fireside cat or dog which only stirs when disturbed, and then methodically finds a comfortable position and goes to sleep again" [10].

The homeostat was tested by placing some of its components under the direct control of the experimenter, by manipulating these components, and by observing the changes in the system as a whole. For example, in a simple situation only two of the four components might be tested [8] Figure 8/4/1. In this kind of study, the feedback being studied was of the type $\mathbf{M_1} \leftrightarrow \mathbf{M_2}$. The relation $\mathbf{M_1} \rightarrow \mathbf{M_2}$ could be placed under the control of the experimenter by manipulating the potentiometer of $\mathbf{M_1}$ by hand instead of using its uniselector. The reverse relationship $\mathbf{M_2} \rightarrow \mathbf{M_1}$ was placed under machine control by allowing the uniselector of $\mathbf{M_2}$ to control its potentiometer. After starting up the homeostat and allowing it to stabilize, Ashby manipulated $\mathbf{M_1}$ to produce instability. The result was one or more advances by the uniselector of $\mathbf{M_2}$, which resulted in stability being re-attained.

Even with this fairly simple pattern of feedback amongst four component devices, many surprising emergent behaviours were observed. For example, in one interesting study Ashby [8] demonstrated that the system was capable of a simple kind of learning. In this experiment, it was decided that one machine ($\mathbf{M_3}$) was to be controlled by the experimenter as a method of "punishing" the homeostat for an incorrect response. In particular, if the needle of $\mathbf{M_1}$ was forced by hand to move in one direction, and the homeostat did not respond by moving the needle of $\mathbf{M_2}$ to move in the opposite direction, then the experimenter would force the needle of $\mathbf{M_3}$ into an extreme position to introduce instability. On the first trial of this study, when the needle of $\mathbf{M_1}$ was moved, the needle of $\mathbf{M_2}$ moved in the same direction. The homeostat was then punished, and uniselector-driven changes ensued. On the next trial, the same behaviour was observed and punished; several more uniselector-driven changes ensued. After these changes had occurred, movement of $\mathbf{M_1}$'s needle resulted in the needle of $\mathbf{M_2}$ moving in the desired direction – the homeostat had learned the correct response. "In general, then, we may identify the behaviour of the animal in 'training' with that of the ultrastable system adapting to another system of fixed characteristics." Ashby went on to demonstrate that the homeostat was also capable of adapting to two different environments that were alternated.

1.1.2 The Tortoise. Ashby's homeostat could be interpreted as supporting the claim that the complexity of the behaviour of whole organisms largely emerges from a) a large number of internal components and from b) the interactions between these components. In the late 1940s, some of the first autonomous robots were built to investigate a counter-claim [10-12]. Grey

Walter's research program "held promise of demonstrating, or at least testing the validity of, the theory that multiplicity of units is not so much responsible for the elaboration of cerebral functions, as the richness of their interconnection" [10]. His goal was to use a very small number of components to create robots that generated much more life-like behaviour than that exhibited by Ashby's homeostat.

Grey Walter (1963) whimsically gave his autonomous robots the biological classification *Machina speculatrix* because of their propensity to explore the environment. Because of their appearance – small tractor-like vehicles surrounded by a plastic shell -- his robots were more generally called tortoises. A very small number of components (two miniature tubes, two relays, two condensers, two motors, and two batteries) were used to create two sense reflexes. One reflex altered the behaviour of the tortoise in response to light. The other reflex altered the behaviour of the tortoise in response to touch.

At a general level, a tortoise was an autonomous motorized tricycle. One motor was used to rotate the two rear wheels forward. The other motor was used to steer the front wheel. The behaviour of these two motors was under the control of two different sensing devices. The first was a photoelectric cell that was mounted on the front of the steering column, and which always pointed in the direction that the front wheel pointed. The other was an electrical contact that served as a touch sensor. This contact was closed whenever the transparent shell that surrounded the rest of the robot encountered an obstacle.

Of a tortoise's two reflexes, the light-sensitive one was the more complex. In low light, the machine was wired in such a way that its rear motor would propel the robot forward while the steering motor slowly turned the front wheel. As a result, the machine could be described as exploring its environment. The purpose of this exploration was to detect light -- when moderate light was detected by the photoelectric cell, the steering motor stopped. As a result, the robot moved forward, approaching the source of the light. However, if the light source were too bright, then the steering motor would be turned on again at twice the speed that was used during the robot's exploration of the environment. As a result, "the creature abruptly sheers away and seeks a more gentle climate. If there is a single light source, the machine circles around it in a complex path of advance and withdrawal" [11].

The touch reflex that was built into a tortoise was wired up in such a way that when it was activated, any signal from the photoelectric cell was ignored. When the tortoise's shell encountered an obstacle, an oscillating signal was generated that rhythmically caused both motors to run at full power, turn off, and to run at full power again. As a result, "all stimuli are ignored and its gait is transformed into a succession of butts, withdrawals and sidesteps until the interference is either pushed aside or circumvented. The oscillations persist for abut a second after the obstacle has been left behind; during this short memory of frustration Elmer darts off and gives the danger area a wide berth" [11].

In spite of their simple design, Grey Walter was able to demonstrate that his robots were very capable of complex and interesting behaviours. He mounted small lights on them, and used long-exposure photography to trace out their trajectories in a fashion that foreshadows Simon's parable of the ant. His records demonstrate that a robot is able to move around an obstacle, and then orbit a light source with complicated movements that do not take it too close, but also do not take it too far away. If presented two light sources, complex choice behaviour is observed: the robot first orbits around one light source, and then wanders away to orbit around the second. If it encountered a mirror, then the light source being used to record its behaviour became a stimulus for its light sensor, and resulted in what became known as the famous "mirror dance". The robot "lingers before a mirror, flickering, twittering and jigging like a clumsy Narcissus. The behaviour of a creature thus engaged with its own reflection is quite specific, and on a purely empirical basis, if it were observed in an animal, might be accepted as evidence of some degree of self-awareness" [10].

1.2 The Synthetic Approach

These two historical examples illustrate two different themes. First, they both show the wisdom of Simon's parable of the ant, in the sense that they demonstrate that complex behaviours can emerge from interactions involving fairly simple components.

Second, they are both prototypical examples of what has become known as the synthetic approach [13]. Most models in classical cognitive science and in experimental psychology are derived from the analysis of existing behavioural measurements. In contrast, both the homeostat and the tortoise involved making some assumptions about primitive capacities, building working systems from these capacities, and then observing the resulting behaviour. In the synthetic approach, model construction *precedes* behavioural analysis.

Braitenberg [13] has argued that psychology should adopt the synthetic approach, because theories that are derived via analysis are inevitably more complicated than is necessary. This is because cognitive scientists and psychologists have a strong tendency to ignore the parable of the ant, and prefer to locate the source of complicated behaviour within the organism, and not within its environment. Pfeifer and Scheier [6] call this the frame-of-reference problem. "We have to distinguish between the perspective of an observer looking at an agent and the perspective of the agent itself. In particular, descriptions of behaviour from an observer's perspective must not be taken as the internal mechanisms underlying the described behaviour".

Here we see one of the strong appeals of adopting the synthetic approach. By building a system and taking advantage of nonlinear interactions (such as feedback between components, and between a system and its environment), relatively simple systems can surprise us, and generate far more complicated behaviour than we might expect. By itself, this demonstrates the reality of the frame-of-reference problem. However, the further appeal of the synthetic approach comes from the belief that if we have constructed the simple system, then we should be in a very good position to propose a simpler explanation of the complicated behaviour. In particular, we should be in a better position than would be the case if we started with the behaviour, and attempted to analyze it in order to understand the workings an agent's internal mechanisms. "Only about 1 in 20 [students] 'gets it' -- that is, the idea of thinking about psychological problems by inventing mechanisms for them and then trying to see what they can and cannot do" (Minksy, 1995, personal communication).

Clearly, the synthetic approach is worth exploring, particularly if it offers the opportunity to produce simple theories of complex, and emergent, behaviours. For this reason, Braitenberg has called for the development of a new approach in psychology that he has named *synthetic psychology* [13]. However, the synthetic approach as it appears in embodied cognitive science is associated with a view that many psychologists would not be comfortable in endorsing.

1.3 Reacting Against Representation

Modern embodied cognitive science can be viewed as a natural evolution of the historical examples that were presented earlier. Researchers have used the synthetic approach to develop systems that generate fascinatingly complicated behaviours [5, 6, 13].

However, much of this research is dramatically anti-representational. "In particular I have advocated situatedness, embodiment, and highly reactive architectures with no reasoning systems, no manipulable representations, no symbols, and totally decentralized computation" [5]. One of the foundational assumptions of behaviour-based robotics is that if a system can sense its environment, then it should be unnecessary for the system to build an internal model of the world.

This is strongly reminiscent of a failed tradition in experimental psychology, called *behaviourism*, that attempted to limit psychological theory to observables (namely, stimuli and responses), and which viewed as unscientific any theories that attempted to describe internal processes that mediated relationships between sensations and actions. "I believe we can write a psychology, define it as Pillsbury, and never go back upon our definition: never use the terms consciousness, mental states, mind, content, introspectively verifiable, imagery, and the like. I believe that we can do it in a few years without running into the absurd terminology of Beer, Bethe, Von Uexküll, Nuel, and that of the so-called objective schools generally. It can be done in terms of stimulus and response, in terms of habit formation, habit integrations and the like" [14].

2. SYNTHETIC PSYCHOLOGY

2.1 The Need For Representation

The resemblance of embodied cognitive science to behaviourism is unfortunate, because it decreases the likelihood that the advantages of the synthetic approach will be explored in psychology. The reason for this is that many higher-order psychological phenomena require an appeal to internal representations in order to be explained.

That stimulus-response reflexes are not sufficient to account for many higher-order psychological phenomena is a theme that has dominated cognitivism's replacement of behaviorism as the dominant theoretical trend in experimental psychology. In the study of language, this theme was central to Chomsky's [15] critical review of Skinner [16]. Many of the modern advances in linguistics were the direct result of Chomsky's proposal that generative grammars provided the representational machinery that mediated regularities in language [17-19]. Similar arguments were made against purely associationist models of memory and thought [20]. For example, Bever, Fodor, and Garrett [21] formalized associationism as a finite state automaton, and demonstrated that such a system was unable to deal with the clausal structure that typifies much of human thought and language. Paivio [22, 23] used the experimental methodologies of the verbal learners to demonstrate that a representational construct – the imageability of concepts – was an enormously powerful predictor of human memory. The famous critique of "old connectionism" by Minsky and Papert [24] could be considered a proof about the limitations of visual systems that do not include mediating representations. These examples, and many more, have lead to the status quo view that representations are fundamental to cognition and perception [1, 2, 25-27].

Some robotics researchers also share this sentiment, although it must be remembered that behavior-based robotics was a reaction against their representational work [5]. Moravec [28] suggests that the type of situatedness that characterizes behavior-based robotics (for example, the simple reflexes that guided Grey Walter's tortoises) probably provides an accurate account of insect intelligence. However, at some point systems built from such components will have at best limited abilities. "It had to be admitted that behavior-based robots did not accomplish complex goals any more reliably than machines with more integrated controllers. Real insects illustrate the problem. The vast majority fail to complete their life cycles, often doomed, like moths trapped by a streetlight, by severe cognitive limitations. Only astronomical egg production ensures that enough offspring survive, by chance". Internal representations are one obvious medium for surpassing such limitations.

The question that this leads to is this: can the synthetic approach be conducted in a way that

	Analytic	Synthetic
Representational	• Production system generated from analysis of verbal protocols • e.g. (Newell & Simon, 1972)	• Multilayer connectionist network for classifying patterns using abstract features • e.g. (Dawson, Boechler & Valsangkar-Smyth, 2000)
Non-Representational	• Mathematical model of associative learning based upon analysis of learning behavior of simple organisms • e.g. (Rescorla & Wagner, 1972)	• Behavior-based robotics system constructed from a core of visuomotor reflexes • e.g. (Brooks, 1989)

Table 1. Classification of research traditions along two orthogonal dimensions: analytic vs. synthetic and representational vs. non-representational

provides the advantages that have been raised in previous chapters, but that also provides insight into representational processing?

2.2 Connectionism, Synthesis, Representation

Of course, the answer to the question that was just raised is a resounding yes. There is nothing in the synthetic approach per se that prevents one from constructing systems that use representations. Describing a model as being synthetic or analytic is using a dimension that it is completely orthogonal to the one used when describing a model as being representational or not. This is illustrated in Table 1, which categorizes some examples of research programs in terms of these two different dimensions.

Synthetic psychology should involve research that is both synthetic and representational. In Table 1, one example of research that fits these two characteristics is connectionist modeling.

With respect to synthesis, connectionist research typically proceeds as follows: First, a researcher identifies a problem of interest, and then translates this problem into some form that can be presented to a connectionist network. Second, the researcher selects a general connectionist architecture, which involves choosing the kind of processing unit, the possible pattern of connectivity, and the learning rule. Third, a network is taught the problem. This usually involves making some additional choices specific to the learning algorithm – choices about how many hidden units to use, how to present the patterns, how often to update the weights, and about the values of a number of parameters that determine how learning proceeds (e.g., the learning rate, the criterion for stopping learning). If all goes according to plan, at the end of the third step the research will have constructed a network that is capable of solving a particular problem.

Many of the early successes in connectionism merely involved showing that a PDP network was capable of accomplishing some task that was traditionally explained by appealing to rule-governed symbol manipulation. However, modern analyses have demonstrated conclusively that a broad variety of PDP architectures have the same computational power as the architectures that have been incorporated into symbolic accounts of cognition [1]. What this means is that a connectionist network can learn to perform any task that can be accomplished by a classical model. As a result, the mere fact that a network can learn a task is no longer an emergent phenomenon of any interest to researchers.

Where, then, does emergence enter a synthetic psychology that uses PDP models? The answer to this question is that while it is neither interesting nor surprising to demonstrate that a network can learn a task of interest, it can be extremely interesting, surprising, and informative to determine what regularities the network exploits. What kinds of regularities in the input patterns has the network discovered? How does it represent these regularities? How are these regularities combined to govern the response of the network? In many instances, the answers to these questions can reveal properties of problems, and schemes for representing these properties, that were completely unexpected. In short, this means that before connectionist modelers can take advantage of the emergent properties of a PDP network that is being used as paradigm for synthetic psychology, the modelers must analyze the internal structure of the networks that they train.

Unfortunately, connectionist researchers freely admit that it is extremely difficult to determine how their networks accomplish the tasks that they have been taught. “If the purpose of simulation modeling is to clarify existing theoretical constructs, connectionism looks like exactly the wrong way to go. Connectionist models do not clarify theoretical ideas, they obscure them” [29].

Difficulties in understanding how a particular connectionist network accomplishes the task that it has been trained to perform has raised serious doubts about the ability of connectionists to provide fruitful theories about cognitive processing. Because of the problems of network interpretation, McCloskey [30] suggested “connectionist networks should not be viewed as theories of human cognitive functions, or as simulations of theories, or even as demonstrations of specific theoretical points”. Fortunately, connectionist researchers are up to this kind of challenge. Several different approaches to interpreting the algorithmic structure of PDP networks have been described in the literature. My students and I have been very successful in generating insights into cognitive functioning by interpreting networks that we have trained on a variety of cognitive tasks.

3. CASE STUDIES

3.1 Spatial Judgements

Dawson, Boechler, and Valsangkar-Smyth [31] trained a particular type of backpropagation network, called a network of value units [32], on psychologically interesting spatial judgement task. Input units were used to represent 13 different cities in Alberta. Output units were used to represent ratings of distance between pairs of cities. The network was trained to make accurate spatial judgements for all possible combinations of city pairs that could be represented.

The hidden units were analyzed by considering them to be analogous to place cells found in the hippocampus [33]. A location for each hidden unit on the map was found that maximized the correlation between connection weights feeding into the unit and distances on the map between cities and the hidden unit location. All of the hidden units could be positioned on the map in such a way that very high correlations between weights and distances were observed.

It was observed that an individual hidden unit's responses to different stimuli were not necessarily accurate. For instance, when presented two cities that were relatively close together, a unit might generate internal activity very similar in value to that generated when presented two other cities that were much further apart. How is it possible for such inaccurate responses to result in accurate outputs from the network?

The answer to this question is that the hidden unit activations in the network are a form of representation called *coarse coding*. In general, coarse coding means that an individual processor is sensitive to a broad range of features, or at least to a broad range of values of an individual feature (e.g., [34]). As a result, individual processors are not particularly useful or accurate feature detectors. However, if different processors have overlapping sensitivities, then their outputs can be pooled, which can result in a highly useful and accurate representation of a specific feature.

Dawson et al. [31] called the representational scheme that they discovered *coarse allocentric coding*. In the literature on the biological foundations of animal navigation, researchers have been very critical of the notion that the hippocampus represents a cognitive map, because single-cell recording studies have shown that it does not exhibit a topgraphically organized, map-like structure. However, the major hypothesis about the hippocampus that was suggested by the spatial judgment network is that place cells also implement a coarse allocentric code. As a result, the place cells need not be organized topographically, because they don't represent the environment in the same way as a graphical map. Instead, locations of landmarks in the environment could be represented as a pattern of activity distributed over a number of different place cells. If this were the case, then in spite of their individual limitations, coarse coding of place cell activities could be used to represent a detailed cognitive map without necessarily being coordinated with other neural subsystems. In other words Dawson et al's [31] discovery of coarse allocentric coding in their network provides one plausible account of how to reconcile the spatial abilities of the hippocampus with its non-maplike organization.

3.2 The Mushroom Problem.

The mushroom problem is a benchmark training set for machine learning [35], and can also be obtained from the UCI Machine Learning Repository. It consists of 8124 different patterns, each defined as a set of 21 different features. The task is to use these features to decide whether a mushroom is edible or not.

Dawson et al. [36] interpreted a network of value units trained a variation of the mushroom problem. This variation involved extra output learning, in which the network not only had to use an output unit to represent whether a mushroom was edible or not, but also had to use other output units to represent the reason for this decision. This network used 21 input units, 5 hidden units, and 10 output units. The first output unit indicated if the mushroom was edible. The remaining nine output units each represented a reason for making a decision, where each reason corresponded to a particular terminal branch in a classical decision tree created for the mushroom problem. The purpose of this network was to determine whether the decision tree could be translated into an ANN using standard connectionist training techniques.

After training, the responses of the 5 hidden units to each of the 8124 patterns were recorded, and k-means cluster analysis was conducted on these responses. It was determined that the patterns of hidden units activities should be assigned to 12 different clusters. Dawson et al. [36] translated the classical decision tree into a set of nine condition-action rules that defined a small production system. They then demonstrated a unique mapping in which all of the patterns that belonged to a particular cluster map directly onto one of these productions. In other words, they were able to show that when the 5 hidden units had a particular pattern of activity -- a pattern that could be assigned to one of the clusters -- this could be translated into a claim that the network was executing a specific production rule. This demonstrates that standard training procedures can be used to translate a symbolic theory into a connectionist network, and blurs the distinction between these two types of theories.

3.3 Implications

The preceding case studies have indicated that one can use connectionism to conduct synthetic psychology, and use the interpretations of networks to contribute to such issues as the debate about the nature of the cognitive map, or the difference between symbolic and PDP models. We have also used this approach to contribute to other psychological domains, including solving logic problems [37], deductive and inductive reasoning [38], cognitive development [39], and the relation between symbolic and subsymbolic theories of mind [40]. Synthetic psychology would appear to be a field that is both tractable and representational.

4. REFERENCES

[1] M. R. W. Dawson, *Understanding Cognitive Science*. Oxford, UK: Blackwell, 1998.

[2] Z. W. Pylyshyn, *Computation And Cognition*. Cambridge, MA.: MIT Press, 1984.

[3] A. Clark, *Microcognition*. Cambridge, MA: MIT Press, 1989.

[4] T. Horgan and J. Tienson, *Connectionism And The Philosophy Of Psychology*. Cambridge, MA: MIT Press, 1996.

[5] R. A. Brooks, *Cambrian Intelligence: The Early History Of The New AI*. Cambridge, MA: MIT Press, 1999.

[6] R. Pfeifer and C. Scheier, *Understanding Intelligence*. Cambridge, MA: MIT Press, 1999.

[7] H. A. Simon, *The Sciences Of The Artificial*, Third ed. Cambridge, MA: MIT Press, 1996.

[8] W. R. Ashby, *Design For A Brain*, Second Edition. New York, NY: John Wiley & Sons, 1960.

[9] W. R. Ashby, *An Introduction To Cybernetics*. London: Chapman & Hall, 1956.

[10] W. Grey Walter, *The Living Brain*. New York, NY: W.W. Norton & Co., 1963.

[11] W. Grey Walter, "An imitation of life," *Scientific American*, vol. 182, pp. 42-45, 1950.

[12] W. Grey Walter, "A machine that learns," *Scientific American*, vol. 184, pp. 60-63, 1951.

[13] V. Braitenberg, *Vehicles: Explorations In Synthetic Psychology*. Cambridge, MA: MIT Press, 1984.

[14] J. B. Watson, "Psychology as the behaviorist views it," *Psychological Review*, vol. 20, pp. 158-177, 1913.

[15] N. Chomsky, "A review of B.F. Skinner's *Verbal Behavior*," *Language*, vol. 35, pp. 26-58, 1959.

[16] B. F. Skinner, *Verbal Behavior*. New York, NY: Appleton-Century-Crofts, 1957.

[17] N. Chomsky, *Aspects Of The Theory Of Syntax*. Cambridge, MA: MIT Press, 1965.

[18] N. Chomsky and M. Halle, *The Sound Pattern Of English*. Cambridge, MA: MIT Press, 1991.

[19] N. Chomsky, *The Minimalist Program*. Cambridge, MA: MIT Press, 1995.

[20] J. R. Anderson and G. H. Bower, *Human Associative Memory*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1973.

[21] T. G. Bever, J. A. Fodor, and M. Garrett, "A formal limitation of associationism," in *Verbal Behavior And General Behavior Theory*, T. R. Dixon and D. L. Horton, Eds.

Englewood Cliffs, NJ: Prentice-Hall, 1968, pp. 582-585.

[22] A. Paivio, "Mental imagery in associative learning and memory," *Psychological review*, vol. 76, pp. 241-263, 1969.

[23] A. Paivio, *Imagery And Verbal Processes*. New York: Holt, Rinehart & Winston, 1971.

[24] M. Minsky and S. Papert, *Perceptrons, 3rd Edition*. Cambridge, MA: MIT Press, 1988.

[25] J. A. Fodor, *The Language Of Thought*. Cambridge, MA: Harvard University Press, 1975.

[26] R. Jackendoff, *Languages Of The Mind*. Cambridge, MA: MIT Press, 1992.

[27] D. Marr, *Vision*. San Francisco, Ca.: W.H. Freeman, 1982.

[28] H. Moravec, *Robot*. New York, NY: Oxford University Press, 1999.

[29] M. Seidenberg, "Connectionist models and cognitive theory," *Psychological science*, vol. 4, pp. 228-235, 1993.

[30] M. McCloskey, "Networks and theories: The place of connectionism in cognitive science," *Psychological science*, vol. 2, pp. 387-395, 1991.

[31] M. R. W. Dawson, P. M. Boechler, and M. Valsangkar-Smyth, "Representing space in a PDP network: Coarse allocentric coding can mediate metric and nonmetric spatial judgements," *Spatial Cognition and Computation*, vol. 2, pp. 181-218, 2000.

[32] M. R. W. Dawson and D. P. Schopflocher, "Modifying the generalized delta rule to train networks of nonmonotonic processors for pattern classification," *Connection Science*, vol. 4, pp. 19-31, 1992.

[33] J. O'Keefe and L. Nadel, *The Hippocampus As A Cognitive Map*. Oxford: Clarendon Press, 1978.

[34] P. S. Churchland and T. J. Sejnowski, *The computational brain*. Cambridge, MA: MIT Press, 1992.

[35] J. S. Schlimmer, "Concept acquisition through representational adjustment," in *Department of Information and Computer Science*. Irvine, CA: University of California Irvine, 1987.

[36] M. R. W. Dawson, D. A. Medler, D. B. McCaughan, L. Willson, and M. Carbonaro, "Using extra output learning to insert a symbolic theory into a connectionist network.," *Minds And Machines*, vol. 10, pp. 171-201, 2000.

[37] M. R. W. Dawson, D. A. Medler, and I. S. N. Berkeley, "PDP networks can provide models that are not mere implementations of classical theories," *Philosophical Psychology*, vol. 10, pp. 25-40, 1997.

[38] J. P. Leighton and M. R. W. Dawson, "A parallel distributed processing model of Wason's selection task," *Cognitive Systems Research*, vol. 2, pp. 207-231, 2001.

[39] C. L. Zimmerman, "A network interpretation approach to the balance scale task," *Unpublished Ph.D. dissertation in Psychology*. Edmonton: University of Alberta, 1999.

[40] M. R. W. Dawson and C. D. Piercey, "On the subsymbolic nature of a PDP architecture that uses a nonmonotonic activation function," *Minds and Machines*, vol. 11, pp. 197-218, 2001.