Suppes FROM STIMULUS-RESPONSE TO BRAIN WAVES ANALYSIS: A TALE ON THE WHITE KNIGHT OF BEHAVIORISM

Patrick Suppes has been called the “white knight of behaviorism”, since some of his early works were dedicated to a hearty defense of a behavioristic approach in psychology. As he has pointed out in more than an occasion, he was not defending a naive behavioristic reductionism à la Skinner, but instead a form of methodological behaviorism, or also neobehaviorism, as Suppes likes to label it: the neobehaviorism retains as main feature the centrality of stimulus-response mechanisms, but does not neglect the fundamental importance of an unobservable internal structure.

This position is completely consistent with Suppes’ ideas on the philosophy of science: the basis of science has to be empirical evidence, and if we are dealing with psychology as a science, observable behavioral evidence in some cases is the only available form of evidence. But this is far from implying that external behavioral responses are all there is to consider when dealing with psychology: Neobehaviorism has been enlightened by the acceptance of a new broader concept of response that includes internal psychological states as a form of response, so much so that we can ask, in the end, for direct or indirect traces of these internal states.

This methodological standpoint in psychology is reflected in the huge amount of work Suppes has done on learning models based on stimulus-response models; his latest Brain Research Project, begun at Stanford in 1996, is the one that really demonstrates how serious Suppes is when talking about internal structure. The goal of the project is to detect invariances among subjects in the production of brain waves concurrent to the performance of language-related tasks: the researchers involved in the project are looking for an internal representation of language, by analyzing the brain wave responses to external stimuli.

This experimental research has so far provided very interesting results, and has led, from a philosophical point of view, to a renewed confidence in the importance of an associationistic approach to psychology. In the article “Invariance of brain-wave representations of simple visual images and their names” (1999) Suppes and colleagues declare that “the results support a quite direct way the solution proposed by Bishop Berkeley and David Hume to a long-standing controversy that began in the 18th century of how the mind represents simple abstract ideas” (p. 14658). The associationistic flavor of this conclusion is not there by chance, and even if Suppes hasn’t written very much about associationism in an explicit and extended way, the importance of this approach in Suppes’ thought is clear from many signs. In this paper I would like to bring together these disparate perspectives, to outline the path that combines Suppes’ behavioristic formation, his associationist convictions, his work on brain representation and his general approach to science. I found these aspects particularly interesting because associationism has been the object of ferocious criticisms in philosophy of mind, as much as behaviorism has; then, in the more recent development in cognitive psychology, the connectionist approach and neural networks models have acquired a very important role. And, getting more familiar with Suppes’ experimental work, it seems to me that his thought has been on this kind of track all the way long; even if he has not been directly involved in the more recent debate, I see how the results of his research represent his best argument.

I begin with a brief overview on behaviorism, focusing mainly on
Suppes' take on it, and in the following section I show some applications of his behavioristic view in his work on learning models; then, in the third section, I jump to his latest work regarding the analysis of brain waves in search of an internal representation of language. In the fourth section I'll try to profile the path that led Suppes from his earliest works to the brain project, which seems connected with his unshaken support of an associationistic approach to psychology.

**BEHAVIORISM**

I feel compelled to spend few words on behaviorism, before getting into the specifics of Suppes' thought about it. Dagfinn Føllesdal, at the beginning of his paper "Intentionality and behaviorism" (1982), has provided one of the best introductions:

*Behaviorism* stands for a variety of attitudes and methodological positions in psychology, from on the one side the epistemological view that in studying man, the sole evidence that our theories can be tested against is observation of his behavior, to on the other extreme, various rather restricted ontological views concerning what man is. In between there are all kinds of positions whose definitions are often so varied and vague that not even their various proponents agree on them, like logical behaviorism, philosophical behaviorism, methodological behaviorism, radical behaviorism, neo-behaviorism, etc. The practitioners are not too much to be blamed for this: "-isms" of all kinds, including for example "positivism", "existentialism", are notoriously difficult to define. As Suppes has pointed out, we have no well-defined framework for such definitions (Føllesdal [1982], p. 553).

Føllesdal is referring here to Suppes' article on behaviorism (Suppes [1969]), where in the opening he expresses a similar skepticism on providing correct definitions or labels for this approach in psychology or philosophy. Instead, coherently with his thought, he attempts a behavioristic description of behaviorism (p. 294):

One initial way to distinguish behaviorism from other approaches to
the study of human beings is in term of the vocabulary used. In behavioristic discussions of human actions or attitudes there continually recur words like 'stimulus', 'conditioning', 'discrimination', and 'reinforcement'. On the other hand, those who are critical of a behaviorist approach, or those who feel it is not adequate to account for all kinds of human behavior, will emphasize such words as 'intention', 'belief', 'purposive behavior', 'rule-following behavior'.

Then, following the linguistic approach to the analysis of behaviorism, Suppes mentions that sentences such as "John believes that there are lions in Alaska" are not just intentional but also intensional: the truth conditions of such belief sentences do not satisfy the ordinary extensional truth-functional logic. This gives him a way to characterize what he thinks behaviorism is for (1969, p. 295):

The widespread and subtle use of intentional sentences in ordinary talk is not something I see any reason for attempting to exercise. The task for the behaviorist presumably is to provide an analysis of the truth conditions for such sentences in nonintentional terms. A large and subtle literature of more than two decades shows clearly enough that this is not a simple or straightforward matter. All the same, I am not at all pessimistic about such an analysis ultimately being given.

From these quotes we can already identify the main characteristics of Suppes' early approach to the matter, and if we compare them with some definitions given for the different branches of behaviorism, Suppes seems to meet at least two. We can read from the Stanford philosophical encyclopedia (Graham [2002]):

Behaviorism is committed in its fullest and most complete sense to the truth of the following three sets of claims.
(1) Psychology is the science of behavior. Psychology is not the science of mind.
(2) Behavior can be described and explained without making reference to mental events or to internal psychological processes. The sources of behavior are external (in the environment), not internal (in the mind).
(3) In the course of theory development in psychology, if, somehow, mental terms or concepts are deployed in describing or explaining behavior, then either (a) these terms or concepts should be eliminated and replaced by behavioral terms or (b) they can and should be trans-
lated or paraphrased into behavioral concepts.

The three sets of claims are logically distinct. Moreover, taken independently, each helps to form a type of behaviorism. "Methodological" behaviorism is committed to the truth of (1). "Psychological" behaviorism is committed to the truth of (2). "Analytical" behaviorism (also known as "philosophical" or "logical" behaviorism) is committed to the truth of the statement in (3) that mental terms or concepts can and should be translated into behavioral concepts.

The first claim could be read as an empiricist statement on psychology: science deals with observable data, and all that is observable in psychology is behavior. And Suppes has always subscribed to empiricism. The other claim that Suppes seems to clearly subscribe to is the third one, connected with the difficulty of evaluation of truth values of statements that include mental terms. As he said, it "is not something I see any reason for attempting to exorcise"; he seems to support this claim only for practical reasons of analysis of language, and not in the attempt to claim what the "correct" vocabulary to talk about psychology is. I think it is worth noting this because behavioristic claims like this are easily taken to mean more than they actually are saying. If we look now at the claim number 2, we can see that its avoidance of mental events is not merely instrumental for scientific investigation or linguistic analysis, it seems to be a denial of any causal effect of mental states. I have not found that Suppes has supported this claim at any time. I think it could be confused with Suppes' empirical approach to learning; it is true that in the nature/nurture debate, someone can argue that Suppes is pending on the nurture side, for instance, in his work on learning systems, he is dedicated to finding the model that can take the most out of experience, minimizing the necessity of referring to preexisting internal structures. But this is far from denying a role to some kind of internal structure.

The vast criticism against behaviorism grew stronger after Noam Chomsky's famous review (1959) of Verbal Behavior by Burrhus F. Skinner. Two strands of Chomsky's criticism seem particularly relevant for our discussion: the importance of internal rules structure and the claim that such rules cannot be the products of learned as-
associations, also known as the argument of poverty of stimulus. Chomsky's argument was focused primarily on language acquisition. Briefly stated, speaking a language is a highly structured performance that seems inevitably directed by internal rules. The investigation of these rules brings up the importance and complexity of mental internal structure, and a hypothesis of innatism is brought up to justify this complexity, because children seem to be able to handle linguistic rules that appear to be radically under-determined by the evidence of verbal behavior offered to the child in the short period in which he or she learns the mother tongue. Both these points, the importance of internal cognitive structure and the innateness hypothesis in explaining learning to compensate the insufficiency of exposure to stimuli, were considered by many the weak points undermining the behaviorist enterprise.

Suppes' position on the matter can be summarized this way: regarding the first point, the idea that behaviorism ignores or even wants to deny internal structure, even if entertained by someone, is in no way a constitutive part of a behavioristic approach. It is true that Skinner's proposal had this radical take, but it is easy to criticize almost every radical stance, and Chomsky's criticism on this point, even if justified in a review of Skinner's work, cannot be seriously considered a criticism of behaviorism on the whole. In later writings, and in particular in his "From behaviorism to neobehaviorism" (1975), Suppes feels compelled to clarify his own viewpoint about the matter, underlining how the importance of internal states has been present in many of his works on learning theories; I will talk more about this in the next section. With respect to poverty of stimulus, Suppes responds with skepticism to the argument that these internal structures must be innate and cannot be formed by interaction with the environment; he says this argument could be converted into a "poverty of innateness" argument: this approach just assumes that all is needed to explain some difficulties in learning theories must be some set of innate rules, and every time you meet a new difficulty, you just put another rule in this black box. To an

---

1 Personal communication, March 2004.
empiricist like Suppes, this seems just the easy way out. Suppes has not been interested in debating this point theoretically; however, the literature\(^2\) that is skeptical about the whole Chomskian enterprise related to innatism and the idea of Universal Grammar\(^3\) is increasing. A promising alternative is the probabilistic constraints approach, mostly related to the use of connectionist models of learning, presented in a clear way, for example, in Seidenberg and MacDonald's (1999) article “A probabilistic approach to language acquisition and processing”. It would be beyond the scope of this paper to illustrate the details of this approach, but it is interesting to note that a probabilistic approach to learning is also very present in Suppes' work on learning models.

**LEARNING MODELS**

One of the earliest and most productive application of behavioristic principles in psychology has been the development of learning theories, starting with the seminal work on classical conditioning conducted by Ivan Pavlov: the investigations undertaken in this framework consider learning as a process by which experience produces some form of change, more or less enduring, in an organism's behavior. Therefore, the fundamental notions to formulate the theory are notions apt to describe behavior in terms of stimulus-response.

The basic idea of a stimulus-response account of learning is simple: an organism can change its behavior through time, and this change seems to be due to its experience in the world. The organism receives stimuli from the world, and as a consequence produces responses. These responses, in turn, result in a positive or negative outcome that has a reinforcing or not reinforcing effect on the organism.

\(^2\) A good starting point would be the collective volume *Rethinking Innatenes* (Elman et al. [1996]).

\(^3\) Universal grammar is defined by Chomsky as "the system of principles, conditions, and rules that are elements or properties of all human languages [...] the essence of human language" (Chomsky [1975]).
with respect to the organism's specific goals or general benefit. These three main concepts, *stimuli*, *response* and *reinforcement*, are the basic ingredients of every stimulus-response system, and this simple recipe was widely used to formulate learning theories in the works of authors like Skinner, Hull and Thorndike. Patrick Suppes' contribution to learning theory started in this tradition, and more specifically his early work on this topic was done side by side with William K. Estes. Estes' classical work (1950), together with Bush and Mosteller (1955) were the earliest attempts to develop formal theories of learning. Estes formulated the *stimulus sampling theory*, while Bush and Mosteller were formulating the framework of *linear model learning theory*, grounded — as well as Estes' framework — in probability theory. Suppes contributed to the systematizing effort in adopting the axiomatic method, and working together with Estes, they achieved a foundational formalization of the two different frameworks, the stimulus sampling and the linear model, allowing a more thorough comparison and fundamental conceptual clarification of the relationship between them. Since this starting, Suppes' work in this field has kept growing. One of his most specific contribution has been in particular that of models for a continuum of responses: in this particular kind of models, the set of responses corresponds to some natural continuum, such as a line segment, so the outcomes of reinforcement are not discrete but correspond to the points on the response continuum. These kinds of models are fit to represent tasks as tracking a movement or steering a vehicle, abilities that are not well described by discrete events.

Even if for many years this kind of work on learning theory was relevant in psychology, at some point there was a significant change in methodology and terminology, that we can relate with the beginning of 'cognitive psychology', marked by Ulrich Neisser book with this title (1967); in this new framework, to which Chomsky gave an important contribution, it was emphasized that the basis for thought were internal rules, and not associations. In name of the primacy of the internal rules, the stimulus-response approach to learning theory was practically abandoned. The main criticism was more or less the same seen before: neglecting the investigation of internal processes
and oversimplification of cognitive tasks such as learning. In this regard, Suppes received a personal criticism in an article by R.J. Nelson [1975]: namely, Nelson pointed out that in Suppes’ article “Stimulus-response theory of finite automata” [1969b], Suppes’ handling of responses does not permit an appropriate concept of internal states, because it requires that all responses have to be observable. Following Nelson’s article, on the same journal, we find Suppes’ reply, “From behaviorism to neobehaviorism”. In this article Suppes refers, first of all, to the difference between determinate and nondeterminate reinforcement (p. 271):

A reinforcement is determinate when the correct response is indicated after the actual response has occurred. For example, if I ask a child the sum of 7+5, then a determinate reinforcement would be the correct answer, 12, when he gave an incorrect answer. An example of nondeterminate reinforcement would be simply to tell him that the answer was incorrect and to ask him to try again.

The first kind of reinforcement was the one used by Suppes in his 1969 theorem; to be able to provide, as reinforcement, the correct answer undoubtedly implies that the response given by the child was observable. Suppes doesn’t deny that, but he wants to underline that he was aware of the limits of such a requirement. He says that he has always been aware that an extension of the model including nondeterminate reinforcement was necessary to give an account of complex learning, and Suppes has actually worked on this in other works. The nondeterminate reinforcement works this way: considering again the child asked to sum 7 and 5, he is not supposed to give the result, but to say if a result presented to him is the right one or not. In this way the child can be told if he’s right or wrong, but the result the child is thinking of doesn’t need to be observable. I have used this example just for symmetry with the previous one, but a better application of nondeterminate reinforcement would be in tasks where the actual response of the subject is not easily articulated. Suppes uses this example: the subject is asked to individuate a set of line drawings generated by a rule, inside a larger set of draw-

---

ings. In every trial, the subject is asked if the drawing shown is part of the special set, and then he is told if he is right or not. In this kind of experiment it is implicit that the answer "Yes" or "No" of the subject is the product of some internal responses, what Suppes calls subtrials, that can be thought of as parts of his hypotheses of which rule is the right one. Reinforcement takes place only after the final response, but not when these other internal responses are given. In this way the model can deal with these internal states, even in eventual cases when internal states could not be expressed and could not be observable. In the light of this, Suppes defines his approach as part of neobehaviorism ([1975], p. 270).

I want to make the essential behavioral feature of neobehaviorism the retention of stimuli and responses as central on one hand, and the introduction of unobservable internal structure as the 'neo' component on the other.

This has been my overview to outline the nature of Suppes' adhesion to behaviorism, underlining that his genuine interest in internal states was limited only by the scarce availability of evidence on such states. This is why in the next section I would like to jump to his latest work, checking how these behavioristic beginnings ended up: we will see that a more significant availability of technology during these years has meant that internal states are no more necessarily unobservable, and huge amount of data are a lot more manageable.

THE BRAIN EXPERIMENTS

During a talk at Stanford, Patrick Suppes explained that his interest in brain research had a chance of becoming a real project only few years after he retired from teaching in 1992; in fact, he ironically remarked, after being retired he found to have some time to kill, and researching the brain seemed like the perfect task to keep himself busy enough. Indeed, he has been working actively since then, officially starting the project in 1996; since then he has gathered around himself several students from different disciplines, he
has published five articles with them, together they are still working on others yet to be published, and a whole chapter of Suppes' latest book, *Representation and Invariance of Scientific Structures* (2002), is dedicated to this brain research.

The main goal of Suppes' brain project is that of collecting data on what is going on in our heads while performing some cognitive tasks, with the hope of tracing invariances in these data that can give us some information about the internal processes and representations. All the experiments conducted so far record electroencephalographic waves (EEG). Scholars have been wondering why they are using EEG when contemporary neuroscience seems predominantly to be done with the technique of functional magnetic resonance (fMRI). The problem is that in Suppes' experiments, the subjects are just asked to look on a screen at a picture or word, or to listen to a word pronounced. The recognition of these stimuli by the subject is very fast, in the order of 300 milliseconds. EEG has a time resolution of at least one millisecond, so it is able to capture what is going on when the brain receives the stimulus; fMRI is a much better technique for localizing brain activity than EEG, but the minimal resolution that it is able to capture is one second, definitely too long for this kind of task.

The basic features of this series of experiments are the following: for every subject, a certain amount of stimuli (for example words) are presented one after the other. Sometimes the same stimulus was presented to the same subject several times. The electric brain waves of the subjects are recorded during every trial (every presentation of the stimulus). Then from all the trials recorded for a certain stimulus (for example, the word 'left'), half of them were averaged to create a prototype for that stimulus, and the other half were averaged to create the test wave for that stimulus. Having reached this stage, the task is to correctly classify the test wave of a random stimulus: the test wave corresponding, for example, to the word 'left' is given as input to a classificatory program; the program compares the test word with all the prototypes waves previously generated from all the stimuli, looking for a match. If it is possible to match the test wave with the prototype one, it means that there is a certain amount of
invariance between different waves recorded for the same stimulus; then we can say we are observing a form of internal representation of the stimulus.

This basic task has been reproduced with variations over several features. The experiments were conducted with different kinds of stimuli (words, sentences, phonemes) presented either visually (written on a computer screen or a picture of an object) or auditorily (through small loudspeakers). The classification task has been performed using prototypes and test waves generated by recordings on a single subject (looking for invariance in the waves of an individual) or, even more significantly, using prototypes and test waves generated by recordings from different subjects. In this case the prototypes for a stimulus generated from a group of subjects were used to classify the test samples for the same stimulus generated from another group of subjects, and a positive result is indicative of invariance in internal representation across different people. Finally, prototypes generated from visual images have been used to classify test samples of brain waves generated by auditory or visual words naming the simple images. These are astonishing results (see Table 1), and if they are further corroborated, they will have for sure an influence on the philosophical debate on representation.

The whole project represents very well several characteristics of Suppes’ method of scientific and, at the same time, philosophical investigation: starting by collecting and analyzing data is faithful to the empirical nature of his approach to science; it exemplifies the actual interplay between experiments and theory formulation, because experiments are designed starting from a theory, but experimental results and availability of better tools drive the reformulation of the theory and the design of more experiments; it makes use of the basic behavioristic notions, providing stimuli and recording responses, but in this case the notion of response being used is the extended neobehavioristic notion of an internal response, that can be considered a sign of internal representation. On top of this, some of the results can be interpreted as a corroboration of the associationist take on representations of abstract ideas. Let’s look at this point more in detail. The specific paper where Suppes advances
this interpretation is the aforementioned Suppes et al. [1999]. In this experiment, some auditory words (like 'square', 'triangle' or 'circle') were presented to the subjects; next, the same words were presented, but printed on the screen; then, the images corresponding to the words, again on the screen. After all this, the prototype brain waves evoked by simple pictures were generated and compared to brain waves evoked by the words naming the pictures. The rate of successful classification was significantly higher than the chance rate, and some pictures in the paper show how the waves for visual images are visibly similar to the ones for auditory or visual words.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Subject</th>
<th># of Successes</th>
<th>Chance Probability</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 visual words¹</td>
<td>S1</td>
<td>32 of 35</td>
<td>1 of 7</td>
<td>91</td>
</tr>
<tr>
<td>7 auditory words¹</td>
<td>S3</td>
<td>34 of 35</td>
<td>1 of 7</td>
<td>97</td>
</tr>
<tr>
<td>12 sentences²</td>
<td>S8</td>
<td>56 of 60</td>
<td>1 of 12</td>
<td>93</td>
</tr>
<tr>
<td>24 visual sentences³</td>
<td>S18</td>
<td>24 of 24</td>
<td>1 of 24</td>
<td>100</td>
</tr>
<tr>
<td>48 visual sentences³</td>
<td>S26</td>
<td>38 of 48</td>
<td>1 of 48</td>
<td>79</td>
</tr>
<tr>
<td>8 visual images⁴</td>
<td>4S</td>
<td>8 of 8</td>
<td>1 of 8</td>
<td>100</td>
</tr>
<tr>
<td>100 visual sentences⁵</td>
<td>S32</td>
<td>88 of 100</td>
<td>1 of 100</td>
<td>88</td>
</tr>
<tr>
<td>100 visual sentences⁵</td>
<td>S32</td>
<td>93 of 100</td>
<td>1 of 100</td>
<td>93</td>
</tr>
</tbody>
</table>

¹ Suppes, Lu and Han [1997]; ² Suppes, Han and Lu [1998]; ³ Suppes, Han, Epelboim and Lu [1999a]; ⁴ Suppes, Han, Epelboim and Lu [1999b]; ⁵ Suppes, Wong, Perreault-Guimaraes, Uy, Yang (to appear).

Table 1 – The first column lists the experiment, described by the prototypes used. The subjects, listed in the second column, are numbered continuously from the experiments first reported in Suppes et al. [1997]. The third column shows the maximum number of test samples successfully recognized out of the total presented. The fourth column shows the chance probability of a correct classification, which is simply 1 divided by the number of prototypes. The fifth column records the percent correct, as computed from the third column.

In the final discussion, Suppes and colleagues underline that these results “strongly support” (p. 14658) the Berkeley-Hume idea that
all general ideas are nothing but particular ones, annexed to a certain term, which gives them a more extensive signification, and makes them recall upon occasion other individuals, which are similar to them (Hume [1739], p. 17).

I think Suppes is also seeing in this a form of confirmation of his fundamental ideas about learning: the internal structures are important, but they are just representations of individual entities encountered during a lifetime experience. What they have in common does not correspond to an internal representation of an abstract concept, but only to a net of associative connections that relates single instances of object—more or less vivid, but most of them connected in a certain degree with the term representing the concept. No other thing is there to represent an abstract concept, but the term that names it, and this term has been, of course, acquired and associated with the rest through experience as well.

I would like to add a personal note about this: independently from Suppes ideas about learning, I must admit that I am skeptic about the conclusion that from these experiments we can definitely confirm Berkeley-Hume idea, because I am not sure that the results trace a picture clear enough of what has been captured. Undoubtedly, the experiments show an invariance between cerebral waves corresponding to the perception of the same stimulus and the perception of the corresponding word, and this invariance must represent something. But it doesn’t seem clear enough to me whether what has been represented is exactly the word. This seems to me the best hypothesis as well, but the evidence doesn’t seem enough to exclude any other hypothesis. There are many alternatives theories on what a concept is and how it is represented in the mind; one of the most popular is, for example, the prototype theory. According to the prototype theory, concepts are represented in the mind by their best or most typical instances, so the concept ‘cat’ would be represented as an image of an average-sized cat with average size, color, hair length and so on. Given this, a supporter of prototypes could claim that what is captured by the brain waves is exactly the repre-

---

sentation of the corresponding prototype, associated to visual stimuli as well as words. Suppes and colleagues bring other arguments in favor of their associationistic conclusion in the paper, based on what we know from other psychological studies on memory. Nevertheless, I think that a more convincing argument could come from other experiments like these, designed to isolate more precisely the source of invariance: I think of more experiments with speakers of other languages, or comparison with blind or deaf subjects, who have different perception modalities available to receive the stimuli, and different vocabulary associated to the stimuli.

EVERYTHING UNDER CONTROL? NOT AT ALL

We have seen that some of the results obtained by Suppes' brain project have been interpreted as a confirmation of one of the basic associationistic ideas. Associationism is a lot more present in Suppes' work than this, and in this section I'm going to explain how. Before getting into the matter, I want to spend few words on what has been perceived as a striking contrast between different characteristics of Suppes' thought.

Some of the features of Patrick Suppes' work have defined him as a rigorous formal thinker: these features are his huge contribution to axiomatic theories, the set-theoretical approach, his efforts in the formalization of foundational concepts of different disciplines, and, coming to psychology, his work of formalization of learning models. From this partial perspective, it's easy to see how someone can think of him as a philosopher devoted to systematicity, whose work during all this time has been driven by a squared conception of the world and a desire for order and control. But it is exactly in his associationistic approach to psychology and philosophy of mind that we can see how far removed he is from this kind of schematic view. This is why I want to stress what is perhaps a lesser known part of Suppes' thought, namely his profound acknowledgment of the messy nature of empirical data, and his rejection of simplistic schematizations of phenomena when we try to use them in more than just an instru-
mental way. Let's see how this view translates into his associationistic ideas on cognitive processes.

When Patrick Suppes talks about cognitive tasks, he doesn't hesitate in using the computational analogy, and he often uses expressions as "to compute the truth of a sentence"; this could lead us to think of him as a hard-core supporter of an identification between the mind and a computer. But when he goes deeper in detail about the nature of such processes, he likes to say that "[Proust] is a better guide to human computation than Turing" (Suppes [2003], p. 142). I've personally seen several of his students being caught by surprise by such a statement. And Suppes is pretty serious about this: all the attempts to describe mind and consciousness as the product of regimented and controlled cognitive processes are unjustified speculations or, in his own words, "fantasies" (Suppes [2003], p. 140): "The actual computations we do are fragmentary, occasional, contextual, driven by associations internal and external".

Associations are indeed a cornerstone of Suppes conception of mind and thought. Suppes often mentions William James and in his Principles of Psychology (1890) as one of the deepest insights into cognitive processes and mental representations, and in his "Rationality, habits and freedom" (2003) he refers to associations as the means of natural computation (p. 146):

The best and hardest mathematical proofs arise not from some linear, nicely formulated line of explicit reason, but from a random, scattered, jumbled associations [...] Only later is an orderly exposition of justification found.

The basic concept of association of ideas, as is well known, started to have an important role in philosophy of mind thanks to the work of British empiricists, first of all David Hume, who formulated some general principles on how associations work. In the 19th century, associationist principles were taken in consideration and carefully reformulated by the pioneers of scientific psychology, foremost among them John Stuart Mill. Here is a very informal description of such principles, taken from John R. Anderson and Gordon H. Bower ([1973], pp. 9-11):
all ideas, and other mental elements, are associated together in the
mind through experience;
- all ideas can be reduced to a basic stock of 'simple ideas';
- these simple ideas are elementary, unstructured sensations;
- simple, additive rules are sufficient to predict the properties of
complex ideas from the properties of the underlying simple ideas.

As is easy to see from these ideas, associationism is a close relative
of behaviorism, and it is fundamental for an empiricist approach to
learning. But associationistic views, as well as behavioristic ones, are
not necessarily committed to deny or neglect internal states. William
James makes use of these principles (except for few elements to
which he does not subscribe) precisely to describe the complexity,
on the whole, of the internal structure; he is not talking about sim-
pel mechanisms of input/output (as stimulus-response mechanisms
are often described). This is even more evident if we explore the lat-
est development of these principles, what has been called (for exam-
ple, by Fodor) neo-associationism, i.e. the connectionist approach.

FROM ASSOCIATIONISM TO CONNECTIONISM?

Edward L. Thorndike (1949), a student of James, introduced the
term "connectionism" in psychology long before the implementa-
tion in artificial neural nets, but since then it has been evident that
the concepts of stimulus and response were more complex and articu-
lated than in many behavioristic models:

For Thorndike, the stimulus was not some punctuate event like the
blink of a light or the tick of a metronome but rather the 'total state
of affairs' perceived by the organism [...] Thorndike's characterization
of 'response' was similarly multifaceted, including not only overt
behaviors but thoughts, interests, and feelings and ranging from ref-
flexes and instincts to learned reactions (Wozniak [1999]).

This deeper articulation is reflected in some of the models real-
ized by artificial neural networks, that make use of distributed rep-
resentations, hidden units, internal feedback from the output units
to the input ones, all features that clearly imply an attention to in-
ternal structure. The connections among units and their activation level constitute the structure that sustains mental activity, and this structure is fundamental in generating future behavior (and not just the immediate response). Without any doubt, connectionist models are often easily interpreted just as stimulus-response mechanisms, but something that is not sufficiently emphasized is that the inputs and the outputs are not always considered to be observable stimuli or responses, they can be thought of as partial results transmitted from a net to another in a systems of nets connected with each other. In this way the interplay between experience and formation of complex mental representations is articulated in a way that cannot be accused to be oversimplified. Going back to Suppes’ work, we have seen before that he wanted to make clear that a good part of his work in learning models is not considering just simple and observable responses (Suppes [1975]), but that he too had a more elaborate conception of what stimuli and responses are. We could also see a similarity between his talking of trials being formed by internal and unobservable sub-trials, and the connectionist talk of representations in neural nets as constituted by sub-symbolic features that are not individually analyzable.

Let’s now take into consideration the main reason why connectionist models are seen as associationistic models. This has to do with the kind of computation going on in the artificial neural nets, to perform learning and reinforcement. It is known that connectionist models are particularly good in extracting regularities from messy data sets, becoming particularly sensitive to patterns of stimuli presented together: this amounts to the same thing as saying that they are really good in recording associations. The more often a set of data is presented together in input to the net, the more the connection between the data inside the representational structure of the net is reinforced. If a connection between events A and B is strong, every time the representation of A is activated, the activation spreads quickly to the representation of B. In other words, A and B are associated. Artificial neural nets are particularly fit to implement this process of spreading activation. The nature of this process is inspired by William James’ description of the “associative law”
(James [1890], p. 534): “When two elementary brain-processes have been active together or in immediate succession, one of them, on reoccurring, tends to propagate its excitement into the other”.

An important character of connectionist models is that they are closely related to probabilistic models. By tuning weighted connections between simple neuron-like units based on experience with events, they pick up on the statistical properties of data. They can actually learn both statistical and structural information at the same time. They have all the features to be considered probability-sensitive learning machines. Even this aspect of connectionism has something in common with Suppes, because we have seen that his work on learning models has been done in the framework of probabilistic learning. And probability theory is on the all very important in Suppes’ thought, we can just think about his contribution in formalization of empirically testable probability theories, and the various works where probability is used for an insight on topics such as inference, causality or decision theory.

Am I trying to claim that the white knight of behaviorism has been, in reality, an *ante litteram* supporter of the connectionist approach? In some way, I would say so, but without stretching this claim too far. If we take into consideration his actual attitude towards connectionism, we can see once again that it is mainly instrumental: in some of his experiments he has been using connectionist models (cf. for example Suppes and Liang [1998]), but just because they proved to be a powerful tool for classification tasks; that is one of the strong points of artificial neural networks. But Suppes has had no real interest in participating to the debate going on about whether connectionist systems are the best candidate to model cognitive processes, in opposition to the classical symbolic systems.

The topic has been raised to popularity by Fodor and Pylyshyn’s famous article “Connectionism and cognitive architecture: A critical analysis”, followed by many replies and counterreplies (cf. MacDonald and MacDonald [1995] for a collection of the main articles). It is important to stress that the debate is not about the

---

6 Cf. Tabor [2003].
formal adequacy of neural networks in performing some kind of computation; it has been demonstrated in many ways that neural networks and Turing machines can calculate the same class of functions. Instead the issue of the debate is about whether a model designed as a classical symbolic system is better than one designed as a neural network in representing mental entities and mental processes. The classical systems explicitly represent mental entities and the rules to be applied to these entities during a cognitive task. With neural networks, we have just one rule, a learning rule, being applied over and over again, and the final result is coming out without insight on a possible structure of the process, or which entities were involved.

Regarding this debate, the only explicit comment I found is in Suppes [2003], where he is saying clearly his opinion, that association processes are the only way to go to model cognitive processes (p. 148), and that he cannot imagine any kind of cognitive task that cannot be expressed by associative models. And he also believes that neural networks are providing a computational structure that is closer, in its organization, to the brain physiology and more effective and appropriate to describe associative processes. But on the other hand Suppes expressed his skepticism on the representational potential of neural networks in a brief commentary to an article by Hanson and Burr (Suppes [1990], p. 507):

[...] from the fact that universal computation, in the sense of equivalence to an universal Turing machine, can be represented by a net, it does not follow that problems of representation are properly understood in any mathematical detail [...] the statistical information given does not get to the heart of the matter; it does not offer a real understanding of how the nets are working.

But, as I said before, it looks like Suppes is not really interested in this kind of debate further than this, at the moment we have only started scratching the surface in the attempt to clarify the organization of the brain. As he reaffirmed several times, at least during lectures and talks, this kind of abstract, theoretical arguments about which is the most appropriate model of brain processes is often an empty quarrel. Only collecting more data from the brain can help us
formulating better models, not speculations on how these models should be. In Suppes' opinion, an important rule in science is to be faithful to data and restrain ourselves from building long arched bridges.

Philosophy Department
Stanford University

REFERENCES

Batchelder and Wexler [1979], "Suppes' Work in the Foundations of Psychology", in Bogan [1979], pp. 149-186.
MacDonald C. and G. MacDonald (eds.) [1995], Connectionism: Debates on Psychological Explanation, Blackwell Publisher Inc., Cambridge (MA).


Suppes P. [1979], "Self-Profile", in Bogan [1979], pp. 3-56.


Suppes P., B. Han, J. Epelboim and Z.L. Lu [1999a], "Invariance of Brain-wave Representations of Simple Visual Images and Their Names", *Proceedings of the National Academy of Sciences* 96: 12953-12958.

Suppes P., B. Han, J. Epelboim and Z.L. Lu [1999b], "Invariance of Brain-wave Representations of Simple Visual Images and Their Names", *Proceedings of the National Academy of Sciences* 96: 14658-14663.


Claudia Arrighi

SUPPES DAI MODELLI STIMOLO-RISPOSTA
ALL'ANALISI DELLE ONDE CEREBRALI:
UN RACCONTO CIRCA IL PALADINO DEL COMPORTAMENTISMO

Riassunto

Una delle aree di ricerca in cui Patrick Suppes ha dato un importante contributo è la psicologia sperimentale, e in questo campo Suppes si è chiaramente schierato, fin dai primi lavori, con l’approccio comportamentista, senza però mai abbracciare, come ribadito in diverse occasioni, un comportamentismo riduzionista a la Skinner. Suppes definisce il suo approccio un comportamentismo metodologico, un neocomportamentismo, che non trascura l’importanza fondamentale di una struttura interna, benché inosserva-bile. In questo articolo, dopo una breve introduzione circa il comportamentismo che si concentra in particolare modo sulla interpretazione che ne dà Suppes, si espongono alcune delle applicazioni di tale approccio comportamentista, in particolare dei modelli di stimolo-risposto, allo studio dei modelli di apprendimento. Da questo si passa ai recentissimi lavori di Suppes dedicati all’analisi delle onde cerebrali alla ricerca di una rappresentazione del linguaggio interna al cervello, delineando il percorso di ricerca che ha portato Suppes dai primissimi lavori di psicologia comportamentista all’attuale progetto di ricerca sulle onde cerebrali, percorso che è connesso strettamente al fondamentale approccio associazionista che Suppes ha sempre adottato nei confronti dello studio della mente.
Patrick Suppes

RESPONSE TO CLAUDIA ARRIGHI

Arrighi has, in her variety of remarks and analyses of my work, caught very well the two most salient features, namely, on the one hand, my interest of long standing in formal axiomatic problems concerning scientific theories, and on the other hand, in sharp contrast my interest in experiments and the detailed analysis of experimental data. Among philosophers I am probably better known for my formal interests than my experimental ones. So in this response I will have a good deal to say about the experimental side of my own thinking and what I believe to be the proper role of the detailed consideration of experiments in the philosophy of science.

Arrighi has a number of different and interesting things to say about behaviorism, some general, some specific to my own views. I will begin with a number of comments about behaviorism. They will very much reflect the changes in my own views as I began to work on the brain in 1996. The first thing to note is how unsatisfactory from a formal standpoint the definitions of behaviorism are – this includes my own, of course. I don’t mean to suggest by this that we should not discuss what we think behaviorism is but rather, that we recognize from the beginning that we will not end up with some satisfactory detailed systematic idea. In this connection the conceptual importance of whether or not to include the brain in discussions of behaviorism is paramount. I am reminded of the contrast

between theories of matter before and after atomism was finally accepted as the correct theory of matter at the beginning of the twentieth century. When I say “correct theory of matter” I mean the correct theory in terms of substantial data available at that time. Moreover for the experiments supporting the existence of atoms, the correctness of the periodic table for the elements, etc. represent experiments that are not false. They stand for all time as a remarkable achievement in the same way as Ptolemy’s astronomy did for 1500 years. They are approximately, correct as much of Ptolemy’s astronomy still is. This doesn’t mean further improvements don’t take place. It’s important, from my standpoint, to have a conception of science, including behavioristic varieties, that have a place for correctness at a certain level of approximation or coarseness, as well as obvious and continued improvements in what had gone before. This seems to be even more true, looking forward, of what we should expect in psychology than it is in physics, where so much has already been accomplished, even though the horizons of physics now seem unbounded in terms of what we can hope to learn over the next several hundred years. Anyway, the introduction in the twentieth century of substantial concentration on the brain has changed psychology for ever. Even though that change has been taking place, – to be conservative about the beginning –, since the excellent discussion of what was known at the time about the brain in the second chapter of William James’ Principles of Psychology (1890) to the modern focus on brain imaging. This comes in four important varieties: electroencephalography (EEG), magnetoencephalography (MEG), positron emission tomography (PET), and functional nuclear magnetic resonance imagery (FMRI) – in the usual biological and medical literature the word nucleus, “nuclear” or an abbreviation for it are omitted because of the fear that medical patients may be disturbed by a form of imaging that refers to nuclear activity. That these methods all represent triumphs of the twentieth century are well known. The lateness of nuclear magnetic resonance is reflected in the fact that the fundamental physics was only fully worked out in the 1940s, for which Bloch and Purcell later received a Nobel Prize. Even MEG only came on the scene in the 1970s be-
cause of the availability of superconducting quantum interference devices to receive the extremely weak magnetic signals from the brain. So brain imaging is something new, and the current torrent of activity is something really new. This doesn’t mean that there weren’t excellent experiments on EEG at a much earlier time, beginning indeed with the early work of Hans Berger starting in 1910.

Well after this review of the growth of brain imaging, the natural question is this. Does the behavior of our brains count as a part of behaviorism? It is certainly true that the classical definitions do not include behavior of our brains, but this is only because at that time so little work had been done, even though some of it in the 1940s and the 1950s was superb. Moreover, various kinds of behavioristic experiments not concerned with the brain were focused on measuring nonobvious responses. Some examples are: Galvanic skin response, heart beat, and careful measurements of the latency of responses, that is, speed of responses, in given situations. So it is natural, it seems to me, now as we get to work in the twenty-first century, to extend behaviorism to include behavior of the brain. The real battle among philosophers will therefore be between those who think that mental activity is just a form of activity of our brains, a naturalistic thesis about the physical nature of the mind as opposed to those who find in the concept of mind something that goes beyond the possibility of a physical account. Obviously I squarely belong to the former, and not the latter, school of thought. Keeping these remarks about the brain in mind, it is clear that many of my own earlier writings about behaviorism need to be modified.

So let me turn to the contrast in the 1969 article on the vocabulary of behaviorism versus that of non-behaviorism. I put in this latter category, as Arrighi quotes, such words as “intention, belief, purpose of behavior, rule-following behavior”. Much more than in 1969, in fact I would say very much more so, I now want to include intention as a concept on the behavioral side. Moreover, I want to start by characterizing, in a way close to Aristotle in the De Anima but with a slightly different flavor, life as animate matter. It is then a feature of animals among living creatures, and therefore a feature of much animate matter, to have intentions. This is not the place to
go into a long argument about why the usual concerns about determinism are misplaced. I have decided views on this and have published them elsewhere (Suppes [1988], [1991], [1993]). Here I'll just take it as a fact that I have an inclusive concept of behaviorism that's happy with intentions and purposes, which, as Aristotle certainly held are just as much a part of the natural world as familiar properties of physical objects. A response to this might be said, well, such a catholic and all-inclusive behaviorism can hardly be rejected by those who believe in mental activity and mental concepts. But my answer is that I am excluding a variety of philosophical thought, namely, all that which is concerned to separate the mental from the physical. A good example would be mental representations. I don't really know what a mental representation is unless what is meant by it is a brain representation, for example, what is a mental representation of the word "isomorphism"? (For more details on my conception of mental representation, see the last section of Chapter 3 of Suppes [2002]).

After that long remark, let me make a minor one about the nature/nurture controversy. It is inevitable, I suppose, that behaviorists are thought to always favor the nurture side of the debate. This is certainly wrong about me and even more so now that I have become much more absorbed in the behavior of the brain and in the transformations of perceptions as they move from the peripheral nervous system to the cortex. Anyone who thought wholly in terms of nurture is surely ignorant about the basic physics of what is going on. Perhaps the thing that is most important, and at the same time most neglected in the discussions of this controversy, is the fact of the enormously complicated physical transformations from every kind of perception essentially into an electrical signal going from the peripheral nervous system to the brain, especially to the cortex. To think for a moment that the intricate machinery of either the auditory or visual system, to take the two most important senses for perceiving language and much else about the world, are derived from scratch as a matter of nurture would be a piece of scientific lunacy. It took millions of years for these two systems to evolve. In fact, rightly put, hundreds of millions years. After this enormously long
period of evolution, it is scarcely surprising that we find it difficult to build robots that have a comparable sense of vision or hearing. On the other hand, one of the great marvels of evolution is how flexible this electrical and chemical system is. It makes itself available for fine tuning the job of nurture not only in the matter of language but the matter of almost all other aspects of interacting with the world. I guess I've come to see the nature/nurture controversy as tedious and uninteresting. The extreme view on either side seems hopelessly wrong.

Much of what Arrighi has to say about learning models in that section of her commentary I agree with. I just want to comment upon how restricted, on the one hand, the concepts of behaviorism are and, on the other hand, how naturally the theory extends into the estimation of parameters for unobservable processes. Let me clarify by two examples. The first is about the rhetoric aimed at stimulus-response theories. In fact, in the mathematical formulations of theories of the 1950s no claim was usually made about the observability of the stimuli taken in by an organism in the experiment. The experimental situation was described, not the stimuli actually perceived by the subjects of the experiment, be they gold fish or humans.

So how was the concept of stimuli dealt with? Well, there was no attempt to have even a theoretical identification of what the stimuli were in all the standard theories that were given formal expression. What was introduced as a parameter was the estimation of the number of stimuli. So the variability in behavior, for example, variation of responses under various partial schemes of reinforcement from one trial to another, would be fit better by the theory by detailed estimation of the number of stimuli.

In a similar vein it was standard to estimate a learning parameter, something that is still done, but that learning parameter was not in itself directly observable, but could only be estimated from theoretical machinery assumed in some particular version of behavioral learning theory given mathematical expression. The only two notions that were left for direct observation were that of response and the reinforcements given under a particular schedule by the experi-
menter. But in detailed theories, even this notion of reinforcement was made subjective in terms of what was perceived by the subjects, as opposed to what had been selected as reinforcement by the experimenter. So the notion of observability was never one that was given rigorous expression in this whole tradition. This is not to be negative about what was done, but just to be realistic from a broad philosophical standpoint about the intricacies of using the concept of observability.

One of the formal things that I concentrated on in learning—theory was the proof that familiar linear models that did not explicitly use a concept of stimulus were the asymptotic limit of stimulus sampling models, where here "asymptotic" means asymptotic in the sense that the number of stimuli approaches infinity. The intricate and detailed proof of this result is to be found in Chapter 8 of Suppes [2002], a proof that is so tedious that my own view of the main interest is in showing how difficult it is to provide absolutely complete reductions of one theory to another even in closely related areas of science. Broader theses of reduction seem hopeless in most cases of serious realization, even though it may be important, as in my own view about the mental as a form of brain activity. This explicit focus represents a fundamental shift for many in the conception of what it means to have a mental life.

I turn now to the brain experiments which have been one of my principal interests since 1996. Let me concentrate on just a few remarks, and try to avoid in the process of doing so, giving an overweening amount of detail about the experiments. First I would like to reformulate what Arrighi has to say about the way I describe the temporal data in our EEG recordings, time-locked to the presentation of a verbal stimulus. From my standpoint the brain is computing first the identification of the transformed character of the stimulus as it moves through either the visual or auditory nervous system to reach the cortex. In this process, first it is transformed into electrical currents that in themselves generate, probably mainly in the synapses of the neurons in the cortex, the electrical field that we record. So our observable data consist of recordings of either an electrical or magnetic field reflecting the electrical activity and secondar-
ily chemical activity of the brain to the reception of such transformed stimuli (I'll speak in terms of electric, I could say electromagnetic or just magnetic but since the experiments are observing the electrical field I shall stick to electrical activity). The point to start with is to emphasize how transformed this electrical activity is in comparison with the physical nature of the stimulus that entered either the rod and cones of the eye or the sound pressure wave that entered the auditory system. Looked at simply from the outside it is utterly remarkable that so much detail in the original source of the stimulus is preserved as invariant. Of course, we are still in the process of discovering these invariants.

My second point is that I am looking then for what in this electrical field that we are recording represents, in the traditional sense of representation, the original verbal stimulus, which itself was a temporal activity. I refer in the case of reading not to the inert printed word on the page, but to the physical activity of observing this word and having electromagnetic phenomenon now at the level of light entering the rods and cones of the eyes. So we are looking for a standard representation, isomorphic in the sense of structure, but remarkably and wonderfully different in external appearance. The difference is almost as great as that between empirical procedures of measurement and isomorphic representations of those procedures to abstract numerical operations on numbers, to give us standard measurements of physical properties or processes.

This return to the image of the representation of concrete experimental procedures of measurement by abstract numerical operations suggests that we could, in principle, aim for the same thing in the case of the brain. So, though it is of course desirable to have a structural isomorphism between spoken speech, for example, and the brain representations of that speech when heard, and also for the even more complicated case when that speech is produced and is initially formulated in the cortex as something to be spoken, so we could go in a different direction and try to make a very detailed science out of the middle, so to speak. Namely, we would produce a detailed abstract structure for what we think of as the mental activities of ourselves as humans, or in simpler cases, of other animals. We
would then want to establish an isomorphism between the brain representation of a spoken sentence and the mental representation of that spoken sentence where the terminology of the brain representation would be physical. In particular it would be electromagnetic in character at the bottom level or at least at the level of EEG recordings. In contrast we would have an abstract formulation in terms of what we would like to think of as the purely mental. Such a program does not seem impossible to work on, but it is quite clear that thinking in philosophy of mind or in cognitive science has in no sense moved very far toward creating such a systematic theory of mental representations. Having such a separate theory of mental representations would in no sense imply we should move away from a completely naturalistic and physicalistic attitude toward the mental, namely, that animate matter can have mental properties, just as it can have electromagnetic properties, mechanical properties, and chemical properties. The viewpoint I am expressing is most certainly not a new one. It is very close to the attitude, expressed in different terminology by Aristotle in the *De Anima* and by Aquinas in his extensive commentary.

Arrighi mentions in several different places and from different angles my insistence on the continuing importance of Humes' central mechanism of the mind, association. I want to say something more here about what appears to be the universal role of association in the brain. Already in the eighteenth century Hume gives a number of good examples of association's role in learning about the world, as well as in the development of the passions, or what we would call now the emotions. Perhaps the cleverest and deepest example of association in the treatise is the analysis at the beginning of Book II of the passion of pride. In the heyday of cognitive science between 1965 and 1980 it was customary to shrug off the mechanism of association as old-fashioned and inadequate to handle the sophisticated concepts being developed by either cognitive psychologists or analytic philosophers. We have by now returned to a sounder view of these matters. The fundamental importance of association in the neurosciences is thoroughly appreciated, and the study of learning processes in terms of association is focused on in both
neuroscience and computer science. A common and mistaken complaint, during those heady days just mentioned, was the claim that, of course, association could not give an account of such complex concepts as those involving rule learning and following. But this is simply a mistake. We now know, as one of the great foundational clarifications of the twentieth century. Rules, or in more technical terminology, any computable function, can be constructed from very simple ingredients, be they a simple Universal Turing Machine’s small number of states and unlimited tape, or an associative network active roles and links. Clear and definite mathematical proofs of these matters are widely available in the literature, so I will not say more. Some clear examples of association in artificial intelligence and machine learning are to be found in papers of my own with others on machine learning of robotic language (Suppes et al. [1995a]; Suppes et al. [1995b]; Suppes and Liang [1996]; Suppes et al. [1966]; and Suppes and Böttner [1998]). The scheme of learning in these machine learning papers is technical and in its own way rather complicated, but is, of course, simplicity itself compared to the much more complicated constructions taking place in our brains. On the other hand, there is much consensus among those working on these matters from a mathematical and a formal standpoint in neuroscience that the two main concepts that need to be understood in terms of how they can be realized in large assemblies of neurons are the concepts of association and of memory storage and retrieval. For a quite recent view of what a realistic neural model looks like to realize these two concepts, see Valiante [2005].

There is one point about Hume connected to a quotation from Bauer and Anderson given by Arrighi that I want to mention. Hume did emphasize that complex ideas were composed out of simple ones. I look upon this as a mistake just as William James did; in fact, it was James’ main dissatisfaction with Hume’s theory of association (James [1890], pp. 594-604). Fortunately, Hume did not give a lot of examples of the sort that would hang him out to dry for several centuries following. He just emphasized simple ideas too much. We could, of course, take another course of simplicity in defense of Hume, but not one that he had in mind, namely, the
development of conditioning in very simple animals, such as *Aplysia*,
is inevitably much simpler than the associations that are the basis of
learning in humans or other higher animals. But this is not a defense
of Hume, just a way of indicating how there is a natural concept of
greater and less simplicity, or, if you prefer the opposite, complexity
in the learning of organisms.

My final comment on Arrighi concerns her closing remarks on
the debate about connectionism and cognitive architecture. Going
back to the well known article of Fodor and Pylyshyn [1988], she
states the alternatives perhaps too simply, but I think she catches the
main point right. My answer is clear. There is extraordinarily weak
positive evidence and much negative evidence that the mental
processing involved in perception or in cognition do not even begin
to approximate the formal systems of inference available in “classical
symbolic systems.” It’s nice to dream that organisms including hu-
mans, are so constructed, with symbols at the ready. There is really
almost no evidence that it is correct or could possibly be correct.
Now, this is not the place to begin the argument all over again. Let
me just remind the reader, however, that the most universal classical
symbolic system is that of classical logic, and it is well known how
little of ordinary reasoning as expressed in ordinary language can in
any direct way be reduced to this formal system. I do not think any-
one who has studied the problem with any carefulness really believes
that, beneath the enormous problems of expressing even the meas-
tured sentences of good lawyers, some simple logical system close to
the classical one is doing the cognitive work. I cannot say positively
exactly how that work is done. Trying to understand in serious de-
tail the physical procedures of natural computation used by the
brain is a worthy endeavor for the future.

REFERENCES

Aquinas T. [ca. 1270/1999], *A Commentary on Aristotle’s De Anima*, Translation by
Robert Pasnau, Yale U.P., New Haven.