A REVISED AGENDA FOR PHILOSOPHY OF MIND (AND BRAIN)

Patrick Suppes
Stanford University
13 December 2006

It is my intention in this article to set forth the case for revision of the topics considered central in the philosophy of mind and to reorient that agenda toward concepts that are more present in current research in psychology and neuroscience. In my view the philosophy of mind should be a subject like the philosophy of physics, dominated in many ways by scientific findings, in this case, in psychology and neuroscience about the nature of the mental. This does not mean there is no room for philosophy left. Rather, its role is changed to the workman-like job of building a conceptual foundation for the scientific results. The building of such a foundation is itself as much scientific as philosophical. What marks it as philosophical is the emphasis on a certain range of concepts, some of which may remain controversial and will not be clarified for some decades by proper theoretical and empirical scientific findings.

I have divided this revised agenda into three parts. The first one deals with computation, perhaps the subject most missing from current philosophical theories of mind. Part two focuses on representation. For reasons that are set forth later, the concentration is on brain rather than mental representations. Finally, the ever present and continually controversial topic of consciousness is looked at anew in the third part, along with habits and their automaticity.

I. Computation

Preliminary remarks on language. A major area of conflict in thinking about mental computation are ideas about language. From the standpoint of much empirical research, which I believe is broadly correct, the computational processes of comprehension or production are almost entirely unconscious in nature. A particularly telling and important example is that of the prosodic features used by a speaker. In so far as these features express anger, fear, contempt, and so forth, they are often more evident to listeners than the speaker, even after they occur. This is a familiar observation, especially about anger.

For many features of speech, we are aware of the results, but not the processes, by which they are produced or comprehended. This will be a general thesis about consciousness discussed later in Part III—awareness of the results, but seldom detailed awareness of the processes themselves.

A radical thesis is that detailed linguistic theories of parsing and generating grammars are mostly wrong as descriptions of brain processing of language. The mental view is that processing is sequential and often conscious in character. In actual fact, most philosophical treatments of the philosophy of mind do not really have a great deal of detailed analysis of this processing. The brain-computation view, in any case, is that processing is massively parallel and mainly unconscious, entirely so in the details. This general thesis I will not defend carefully at this point, but will amplify as I turn to more concrete questions.

What I said about parsing applies also to determining meaning or, what I prefer, truth. The folklore account is that we parse the sentence and then determine its meaning.

The brain-computation view is that the processing happens nearly simultaneously and in parallel. We have no awareness at all of how we compute the truth value, just as we do not have awareness of how we compute the grammatical correctness of an utterance. So, if I give persons sentences they have not heard before about the geography of Europe, they will be able to tell me rather quickly, within about a 150 milliseconds after the end of each sentence, whether it is true or false. Typical examples are *Paris is not the capital of Poland* and *Rome is north of London*.

I return to these broad remarks later in talking in a more detailed way about computation, but first I want to turn to some philosophical views of the mental representation of language and their relation to computation.

Mental representations. Here is a quote from Jonathan Lear about Quine and Wittgenstein with which to begin,

> Quine, like Wittgenstein, categorically rejects the notion that meaning can essentially involve anything private to an individual, such as a hidden mental image. This is the myth of the museum—that words name specimen mental objects—which Quine rightly urges us to reject. If we are to explain language-mastery, and thus the meaning our words and sentences have, we must do it on the basis of our experience: the sensory evidence of all types to which we have over time been exposed and our sensitivity to it. Positing interior mental objects that are named by words only gets in the way of an explanation, for it merely papers over the gaps in our understanding of how language-mastery is acquired.
>
> (Lear, *Going Native*, pp. 177–78)

Now let us turn to the contrasting view of Noam Chomsky.

> As I am using the term, knowledge may be unconscious and not accessible to consciousness. It may be "implicit" or "tacit." No amount of introspection could tell us that we know, or cognize, or use certain rules or principles of grammar, or that use of language involves mental representations formed by these rules and principles. We have no privileged access to such rules and representations. This conclusion appears to be inconsistent with requirements that have been widely enunciated. Kant, for example, insisted that "All representations have a necessary relation to a *possible* empirical consciousness. For if they did not have this, and if it were altogether impossible to become conscious of them, this would practically amount to the admission of their non-existence." Similar doctrines are familiar in contemporary philosophy. John Searle writes that "It is in general characteristic of attributions of unconscious mental states that the attribution presupposes that the state can become conscious, and without this presupposition the attributions lose much of their explanatory power."
>
> (Chomsky, *Rules and Representations*, p. 128)

My summary view is that, in historical order, Kant, Wittgenstein, Quine, Searle, and Lear are wrong. Chomsky is right. Our conscious mental representations of language are poorly developed, especially almost all aspects of processing, i.e., computation. As in other areas of the mind, we are conscious of results not mental processes, a point I have stressed earlier and will do so later.

On both sides that I have quoted, what is missing, of course, are detailed answers. The views of Lear are particularly mysterious. To turn back on him his own concept of mystery, what in the world is his theory of how language is processed? Surely, there is something going on in the brain. It is not just a matter of external sensory experience. The absence of any conjectures whatsoever about how syntactic and semantic computations are made by the brain is the most striking feature, in fact, of what Wittgenstein and Lear have to say. Quine, who is more sensible than either of them, hints in various accounts of stimulus meaning and the like at the psychology, if not the neuroscience, of language learning. Chomsky also has detailed things to say about what the rules must be like and he is in favor, generally, of a biological view toward language. It is just that he has not ventured into the psychology or neuroscience of such matters in much depth.

Problem of the computation of truth. I turn now to an important specific case of computation. One of the scandals of both philosophy and linguistics, as well as psychology, is the absence of any detailed theory of how the most elementary empirical sentences are judged true or false, or to put it more directly, how their truth value is computed. Consider atomic sentences. Tarski's semantic theory of truth offers no help in determining their truth. He was not concerned to give a theory of how we compute or know that individual atomic sentences such as *Paris is north of Rome* are true or false. Computation of their truth is not generally familiar; I have introduced the subject for deliberate, and indeed necessary, purposes here. There are only one or two views practically possible. The truth of falsity of such sentences is computed from the knowledge and beliefs an individual is able to construct—and I say ‘construct’ not ‘store’—to emphasize processing, or the truth value is simply a mysterious act of the mind beyond comprehension. Obviously, the latter view is absurd. But how are such elementary sentences computed? There are the outlines of a theory in parts of psychology. I shall develop these ideas here without entering into extensive technical details.

It is possible that some, perhaps even many, philosophers who read what I have just written will think they remember that the problem of the computation of truth actually was solved in a nearly satisfactory way by Tarski in his famous article of 1936 “The concept of truth in formalized languages”. Everyone who may have forgotten the details will still remember the famous criterion of truth illustrated by: ‘*it is snowing*’ is a true sentence if and only if it is snowing, which, as Tarski (1936/1983, p. 156) remarks, we are assuming is the case. On the preceding page Tarski states a familiar criterion that, as he says, is more or less that of the classical view of the truth, “a true sentence is one which says that the state of affairs is so and so, and the state of affairs indeed is so and so”. In this connection, he quotes also the earlier source of a very similar remark, well known in philosophy, of Aristotle in the *Metaphysics* (1011b27), “to say of what is that it is or of what is not that it is not, is true”. Intuitively put, Tarski's task is not at all to investigate what is the state of affairs. His concern is with the recursive definition of truth, when the truth of atomic sentences is given. So, if the truth of atomic sentences is

known, we can then compute the truth of any complicated sentence recursively from that of the atomic sentences. What I am saying is not meant to be a technical characterization of Tarski's recursive definition, which explicitly holds only for certain formal languages, but rather to give a feeling for it. It is the truth value of the empirical atomic sentences that is in no sense touched by his characterization, and he understood this perfectly well. It is also not my task here to actually investigate the empirical truth of some indefinite number of atomic sentences. It is rather different. It is how we use our beliefs, our memories, and our associations of other kinds to compute, from this basis, the truth of many different empirical atomic sentences about familiar phenomena. Certain ones we are expected to compute easily, even if we have never heard them uttered before. Such simple geographic examples are what I consider here. We could just as well pick familiar political events of the last twenty-five years. We would also expect to get very quick answers, without any reference to outside sources of knowledge. So, as you can see, what I want to turn to is the psychological theory of how the mind, i.e., brain, of a person computes the truth of new sentences, when the truth should be evident from things already known by that person.

Preliminary remarks on association. The investigation is simply of the brain computations, or if you wish, mental computations, needed to compute the truth value of a sentence that has probably not been previously heard or read. A more detailed formulation of ideas is given in Suppes and Béziau (2004), but the supporting neuroscientific literature is not cited there either, and many gaps remain in our scientific understanding of the neural processing.

The general ideas I use are not new. They derive from Aristotle and Hume, the Godfathers of what I consider a sound philosophy of the mind. In the direct matter of computation I will refer to Hume, and delay consideration of Aristotle until the next section on representation. Psychological and neural methods of computation are the focus. I will use both kinds of description, because for me they are one and the same. What is psychological is embodied in the brain. What is claimed to be psychological and not embodied in the brain is only fantasy. But often we have not yet discovered exactly how a given psychological process is realized in the brain, and so we have to leave open the neural details.

Hume was the first to state so absolutely clearly and unequivocally that there are really just three basic mechanisms of computation in the mind—of course, he did not use the word ‘computation’. The classic formulation is to be found in Section 4 of Book I of the *Treatise of Human Nature* (1739/1951). He says at the beginning that the faculty of imagination, must be guided by some universal principles. These are the mechanisms of association.

> The qualities, from which this association arises, and by which the mind is after this manner convey'd from one idea to another, are three, viz. RESEMBLANCE, CONTIGUITY in time or place, and CAUSE and EFFECT.
>
> (Hume, *Treatise*, p. 11)

What is remarkable about this passage is that not only does Hume put his finger on the principal role of association, he also has an excellent hypothesis of what are the main associative mechanisms. One perception of an apple seeming so much like another gives

rise to forming the idea of an apple; constant contiguity between fire and smoke, leads to a causal association. Hume goes on to say that of the three, the recognition of cause and effect, that is, that from a certain occurrence there follows another event, which we recognize as that of effect, is the most extensive, and in many ways the most important. I emphasize this remark about cause and effect, for it is sometimes mistakenly thought that Hume was a skeptic about causal notions, because he denied, in contradistinction to Locke and Newton, their necessary character. He was also rightly constrained in recognizing that it is not possible to go beyond a certain depth without additional kinds of knowledge that he did not have. Here is the quotation on this topic, that I like the most. He properly characterizes it as part of the nature of a true philosopher to restrain the intemperate desire to continue searching for causes when he is only led into `"obscure and uncertain speculations".

> These are therefore the principles of union or cohesion among our simple ideas, and in the imagination supply the place of that inseparable connexion, by which they are united in our memory. Here is a kind of ATTRACTION, which in the mental world will be found to have as extraordinary effects as in the natural, and to shew itself in as many and various forms. Its effects are every where conspicuous; but as to its causes, they are mostly unknown, and must be resolv'd into *original* qualities of human nature, which I pretend not to explain. Nothing is more requisite for a true philosopher, than to restrain the intemperate desire of searching into causes, and having establish'd any doctrine upon a sufficient number of experiments, rest contented with that, when he sees a farther examination would lead him into obscure and uncertain speculations.
>
> (Hume, *Treatise*, p. 12--13)

It is also important to catch the reference to attraction, because Hume felt, in many ways properly, that the role of association in the life of the mind was as dominate and as significant as that of gravitation in the motion of physical objects.

The first edition of Hume's *Treatise* was published in 1739. So now more than 250 years later, we have more elaborate things to say, but in spite of qualms about association in various intervening periods, it has now returned to its proper place in the theory of the brain. Well noticed is the large association area of the human cortex, in comparison even to other primates. Associative networks are being developed by scientists from many disciplines, and for a variety of purposes. Mistakenly, some philosophers and psychologists, only a few decades ago, would have scoffed at the idea that any serious computations of any kind could be made by principles of association. But this is simply a conceptual mistake, well recognized in the general theory of computation. There are now a number of papers showing the computational power of associative networks and it is easy to prove that we can simulate with an associative network a universal Turing machine capable of computing any computable, i.e., any partial recursive, function, in the standard theory of computation. This universal Turing machine need in principle not be large. One of the best examples, even after a good many years, is the one introduced by Marvin Minsky (1967) consisting of four output symbols and seven internal states. So that using the commonly applied measure of the product of the number of symbols and

the number of states, Minsky's universal Turing machine has the number 4 x 7 = 28. Not quite the best, but almost as good as any that can be reached according to present arguments and examples. I will not go through the details of how such a Turing machine can be simulated by an associative network. It is a familiar fact of modern computational theory that rather simple devices, including Minsky's machine, are quite adequate for computing anything in principle. This does not mean they would serve as practical computers. In like fashion, it is quite another matter to have a deep and detailed understanding of the computations actually made by the brain.

My next task here is much simpler. I propose, schematically, how we can, using psychological and neural concepts, compute from associations the truth of obvious empirical sentences.

Associative networks. The remarks thus far are a prolegomena to giving a sketch of the theory of how such ordinary computations of truth are made. The basic idea is that the computations are made by an associative network, with brain representations of words being the nodes and the links between the nodes being the associations. More generally, auditory, visual, and other kinds of brain images can also be nodes. There is a reasonable body of evidence to support the hypothesis that the nodes of the network are collections of synchronized neurons.

In the initial state, not all nodes are linked, and there are, in this simple formulation, just two states, *quiescent* and *active*. No learning or forgetting is considered. It is assumed that, after a given sentence is responded to as being either true or false, all the activated states return to quiescent. The axioms, which are not stated here, are formulated just for the evaluation of a single sentence, not for giving an account of how the process works over a longer slice of discourse. The way to think about the networks introduced is that a person is asked to say whether a sentence about familiar phenomena is true or false. It is very natural to ask, and not to have a quibble about 'Do you believe this, even though you don't know whether it is true?'

The sentence input comes from outside the associative network in the brain. I will consider only spoken words forming a sentence, although what is said also applies to visual presentation, as well. So, as the sentence is spoken, the sound-pressure image of each word that comes to the ear is drastically transformed by a sequence of auditory computations leading to the auditory nerve fibers, which send electromagnetic signals to the cortex. Such signals are examples of those mentioned earlier. In previous work, I have been much concerned with seeing if we can identify such brain signals as brain representations of words. Some references are Suppes, Lu, and Han (1997) and Suppes, Han, Epelboim, and Lu (1999a, 1999b).

The brain activates quiescent states by using the signal brought into the cortex as the brain representation of the verbal stimulus input. With the activation of the brain representation of words by external stimuli, the associations between activated brain representations are also activated by using this same signal.

Moreover, it is assumed in the theory that activation can be passed along from one associated node to another by a phenomenon characterized some decades ago in psychological research as *spreading activation*. For example, in a sentence about a city like Rome or Paris, some familiar properties are closely associated with these cities and the brain representation of these properties may well be activated shortly after the activation of the brain representations of the words *Rome* or *Paris*, even though the

names of these properties, or verbal descriptions of them, did not occur in any current utterance. This is what goes under the heading of *spreading activation*. Some form of it is essential to activate the nodes and links needed in judging truth, for, often, we must depend upon a search for properties, which means, in terms of processing, a search for brain representations of properties, to settle a question of truth or falsity. A good instance of this, to be seen in the one example considered here, is the one-one property, characteristic of being a capital: *x is capital of y*, where *x* is ordinarily a city and *y* a country. There are some exceptions to this being one-one, but they are quite rare and, in ordinary discourse, the one-one property is automatically assumed. But this is only one of many examples, easily given, that arise in ordinary conversation. (For computer-science applications of spreading activation to information retrieval, see Crestani (1997) for a critical survey.)

One other notion introduced in the axioms of Suppes and Béziau (2004) for computing truth is the concept of the *associative core* of a sentence, in our notation, *c(S)* of a sentence *S*. For example, in the kinds of geography sentences given in the experiments referenced above, where similar syntactic forms are given and the sentences are given about every four seconds, persons apparently quickly learn to focus only on the key reference words, which vary in an otherwise fixed sentential context, or occur in a small number of such contexts. So, for example, the associative core of the sentence *Berlin is the capital of Germany* is a string of brain representations of the three words *Berlin*, *capital* and *Germany*, for which I use the notation BERLIN/CAPITAL/GERMANY, with, obviously, the capitalized words being used to denote the brain representations. A more complicated concept is needed for more general use.

In the initial state of the network, associations are all quiescent, e.g., PARIS ~ CAPITAL, and, after activation, we use the notation PARIS $\approx$ CAPITAL. In the example itself, we show only the activated associations and the activated nodes of the network, which are brain representations of words, visual or auditory images, and so forth. The steps of the associative computation are numbered in temporal steps t_1, etc., which are meant to include some parallel processing.

Example. *Rome is the capital of France*.

t_1.	ROME, CAPITAL, FRANCE	Activation
t_2.	PARIS, 1–1 Property	Spreading activation
t_3.	ROME $\approx$ CAPITAL, CAPITAL $\approx$ 1–1 Property CAPITAL $\approx$ FRANCE, PARIS $\approx$ CAPITAL PARIS $\approx$ FRANCE	Activation
t_4.	ITALY	Spreading activation
t_5.	PARIS/CAPITAL/FRANCE ROME/CAPITAL/ITALY	Activation
t_6.	TRUE $\approx$ PARIS/CAPITAL/ FRANCE TRUE $\approx$ ROME/CAPITAL/ITALY	Spreading activation
t_7.	FALSE $\approx$ ROME/CAPITAL/FRANCE	Spreading activation

This sketch of an example, without stating the axioms and providing other technical details, is meant only to provide a limited intuitive sense of how the theory can be developed for simple empirical sentences. Most important, there is no account of learning associations. Only an idealized performance setup is considered.

II. Representation

A central topic in the philosophy of science is the analysis of the structure of scientific theories. Much of my own work has been concerned with this topic, but in a particular guise. The fundamental approach I have advocated for a good many years is the analysis of the structure of a theory in terms of the models of the theory. In a general way, the best insight into the structure of a complex theory is by seeking representation theorems for its models, for the syntactic structure of a complex theory ordinarily offers little insight into its nature.

In attempting to characterize the models of a theory, the notion of isomorphism enters in a central way. Perhaps the best and strongest characterization of such models is expressed in terms of a *significant representation theorem*. By such a theorem for a theory the following is meant. A certain class of models of the theory, distinguished for some intuitively clear conceptual reason, is shown to exemplify within isomorphism every model of the theory. More precisely, let $\mathfrak{M}$ be the set of all models of a theory, and let $\mathfrak{B}$ be some distinguished subset of $\mathfrak{M}$. A representation theorem for $\mathfrak{M}$ with respect to $\mathfrak{B}$ would consist of the assertion that given any model M in $\mathfrak{M}$ there exists a model in $\mathfrak{B}$ isomorphic to M. In other words, from the standpoint of the theory every possible variation of model is exemplified within the restricted set $\mathfrak{B}$. It should be apparent that a trivial theorem can always be proved by taking $\mathfrak{B} = \mathfrak{M}$. A representation theorem is just

as interesting as the intuitive significance of the class $\mathfrak{B}$ of models, and no more so. An example of a simple and beautiful representation theorem is Cayley's theorem that every group is isomorphic to a group of transformations. One source of the concept of a group, as it arose in the nineteenth century, comes from consideration of the one-one functions which map a set onto itself. Such functions are usually called transformations. It is interesting and surprising that the elementary axioms for groups are sufficient to characterize transformations in this abstract sense, namely, in the sense that any model of the axioms, i.e., any group, can be shown to be isomorphic to a group of transformations.

Philosophical views of mental representations. Without attempting anything like a detailed and accurate account of the long and complicated history of the concept of mental representation in philosophy, it can be enlightening and relevant to review some of the standard conceptions and issues from Aristotle onward. In fact, here I mainly restrict myself to the Aristotelian and Humean traditions. (For more details, see Suppes (2002, Ch. 3).) The main point will not be to assess the correctness or to criticize the adequacy of the analysis given, but to reflect on whether or not there is a notion of isomorphism in the background, as reflected in such concepts as likeness or resemblance. Comment will also be made as to how such notions are tied to representations.

Aristotle. That the defining feature of sense perception is receiving the form of a physical object is, in general terms, the view of perception for which Aristotle is most famous. As he works out the details, it is not an unreasonable view at all, even though it is quite obvious from a modern viewpoint that it cannot be entirely correct.

The background of Aristotle's discussion of perception and implicitly, therefore, of mental representation, is his distinction between form and matter, which applies to objects and phenomena of all kinds. For example, the form of an axe is different from the matter that receives that form. This distinction is also relative in the sense that, for example, the matter of a house can be made up of bricks, which, in turn, have their own form and matter, of which they are made up. I defended the continued viability of this general concept of matter in Suppes (1974).

Aristotle makes important use of the distinction between form and matter in his theory of perception. It is the role of the sense organ to have the potential to receive the form, but not the matter, of a physical object, as described in this passage from the second book of the *De Anima*:

> We must understand as true generally of every sense (1) that sense is that which is receptive of the form of sensible objects without the matter, just as the wax receives the impression of the signet-ring without the iron or the gold, and receives the impression of the gold or bronze, but not as gold or bronze; so in every case sense is affected by that which has colour, or flavour, or sound, but by it, not *qua* having a particular identity, but *qua* having a certain quality, and in virtue of its formula; (2) the sense organ in its primary meaning is that in which this potentiality lies.
>
> (Aristotle, *De Anima*, 424a17–424a25)}

So, if in perceiving the candle we receive exactly its form, but not its matter, then we have a representation of the physical candle that is isomorphic to it in the mind, just because of the sameness of form. The relevant point here is that this notion of the

sameness of form corresponds very closely to the concept of isomorphism that I have been arguing for as a central concept of representation. The concept of form catches nicely that of isomorphism, in the sense that the form does not thereby refer to all the properties of the candle, but only to the properties perceived by the various senses of sight, touch, etc. This kind of restriction of the properties considered is, as already noted in earlier discussions, characteristic of any workable notion of isomorphism.

Upon reading the above passage and my comments, someone might reflect that this is not a very rich theory of the mental, but only of perception. Aristotle, however, goes on to extend the same ideas about forms to the intellect. It is not possible here to discuss all the subtleties involved in his worked-out theory, but the following passage makes clear how the part of the soul that thinks and judges operates in the same way with forms and, thus, the characteristic notion of isomorphism is again applicable:

> Concerning that part of the soul (whether it is separable in extended space, or only in thought) with which the soul knows and thinks, we have to consider what is its distinguishing characteristic, and how thinking comes about. If it is analogous to perceiving, it must be either a process in which the soul is acted upon by what is thinkable, or something else of a similar kind. This part, then, must (although impassive) be receptive of the form of an object, *i.e.*, must be potentially the same as its object, although not identical with it: as the sensitive is to the sensible, so must mind be to the thinkable. …Hence the mind, too, can have no characteristic except its capacity to receive. That part of the soul, then, which we call mind (by mind I mean that part by which the soul thinks and forms judgments) has no actual existence until it thinks. So it is unreasonable to suppose that it is mixed with the body; for in that case it would become somehow qualitative, *e.g.*, hot or cold, or would even have some organ, as the sensitive faculty has; but in fact it has none. It has been well said that the soul is the place of forms, except that this does not apply to the soul as a whole, but only in its thinking capacity, and the forms occupy it not actually but only potentially.
>
> (Aristotle, *De Anima*, 429a10-18, a23-30)

It is a definite part of Aristotelian thought that the forms do not exist separate from individual bodies. There is no separate Platonic universe of forms. A very clear statement on this Aristotelian view is made by Aquinas in the following passage.

> I say this, because some held that the form alone belongs to the species, while matter is part of the individual, and not of the species. This cannot be true, for to the nature of the species belongs what the definition signifies, and in natural things the definition does not signify the form only, but the form and the matter. Hence, in natural things the matter is part of the species; not, indeed, signate matter, which is the principle of individuation, but common matter. For just as it belongs to the nature of this particular man to be composed of this soul, of this flesh, and of these bones, so it belongs to the nature of man to be composed of soul, flesh,

> and bones; for whatever belongs in common to the substance of all the individuals contained under a given species must belong also to the substance of the species.
>
> (Aquinas, *Summa Theologica*, I, Q.75. Art. 4)

The discussion and citations given do not adequately portray the rich tradition of Aristotelian psychology. Aristotle's *De Anima* and his many commentators, including important Arabic ones, but especially Aquinas, established a long intellectual tradition still vigorously reflected in such seventeenth-century works as Descartes' *Passions of the Soul* (1649/1927). It has not been sufficiently remarked that this tradition of Aristotelian psychology, to label it in modern terms, is in its own way comparable to the glorious Ptolemaic tradition in astronomy, which only ended with the *Astronomia Nova* (1609) of Kepler.

Such clear and detailed expositions of the *De Anima* as Themistius' "Paraphrase" (1996) written in late antiquity (fourth century) were still widely read, in the Renaissance. Indeed, Aquinas' commentary formed the basis of Thomistic psychology, still actively studied as a systematic subject in the twentieth century (Brennan, 1941).

Hume. On various grounds, in the *Treatise of Human Nature*, Hume argues vigorously against any direct concept of mental representation between objects in the external world and what is in the mind. He has many things to say about these matters.

The first point to emphasize has already been remarked on in the preceding section. The mechanism of the mind for Hume is association. The important point relevant to the discussion here is the pride of place that Hume gives to resemblance, since only it is a quality, as he puts it, that holds between impressions. And, of course, resemblance is exactly Hume's notion close to that of the modern notion of isomorphism as a basis of representation.

> …'Tis plain, that in the course of our thinking, and in the constant revolution of our ideas, our imagination runs easily from one idea to any other that *resembles* it, and that this quality alone is to the fancy a sufficient bond and association.
>
> (Hume, *Treatise*, p.~11)

I said earlier that Hume denies a direct resemblance between fragmentary perceptions and the real objects posited to generate those perceptions, but he does defend a realistic notion of physical objects and their continued identity in time. His long and complicated argument about these matters is one of the most substantial parts of Book I of his *Treatise*. I want to point out various ways in which he uses the concept of resemblance or isomorphism. The following passage is particularly striking, because he argues for the constancy of our perception of something like the sun or other very stable objects by using the fact that our successive perceptions, though interrupted, resemble each other and justify, therefore, as he argues in detail, the inference to identity across time of the object. Here is the key passage about this use of resemblance.

> When we have been accustom'd to observe a constancy in certain impressions, and have found, that the perception of the sun or ocean, for

> instance, returns upon us after an absence or annihilation with like parts and in a like order, as at its first appearance, we are not apt to regard these interrupted perceptions as different, (which they really are) but on the contrary consider them as individually the same, upon account of their resemblance. But as this interruption of their existence is contrary to their perfect identity, and makes us regard the first impression as annihilated, and the second as newly created, we find ourselves somewhat at a loss, and are involv'd in a kind of contradiction. In order to free ourselves from this difficulty, we disguise, as much as possible, the interruption, or rather remove it entirely, by supposing that these interrupted perceptions are connected by a real existence, of which we are insensible. This supposition, or idea of continu'd existence, acquires a force and vivacity from the memory of these broken impressions, and from that propensity, which they give us, to suppose them the same; and according to the precedent reasoning, the very essence of belief consists in the force and vivacity of the conception.
>
> (Hume, *Treatise*, p. 199)

Hume does not stop here. Over several pages he explains four different aspects of this setup in the inference from interrupted perceptions to identity of objects through time. It is one of the most sustained arguments in the *Treatise* and a wonderful piece of philosophical analysis. He first explains the basis for the principle of the individuation of objects, and he introduces here something that is very much in the spirit of isomorphism and a relevant notion of invariance.

Hume's ideas dominated much of the psychology of the nineteenth century, to which even William James agreed in his influential treatise *Principles of Psychology* (1890). James was quite critical of Hume's use of simple ideas, out of which to build complex ones. He objected to such as atomistic concept as central to the nature of thought.

Brain representations of language: general remarks. When a person hears the word *Paris* or reads the word *Italy*, usually in a sentential context, the encoded word token reaching the cortex must be recognized, in some way or another, as isomorphic to a word already encoded in long-term memory. Of course, none of this process of recognition is conscious in the usual settings in which we listen or read. Given that this description is roughly correct, a first problem of research is to find, if possible, brain-wave representations of the initial processing of verbal stimuli. There is a substantial tradition of studying components of these waves, usually called evoked response or event-related potential (ERP) components, generated by various verbal stimuli, e.g., visually presented sentences with semantic anomalies. For reviews of this literature, see Brown and Hagoort (1999), and Rugg and Coles (1995). The task considered here is different, namely, identifying particular waves as the token representations in the brain of auditorily or visually presented words or sentences. From a theoretical or statistical standpoint, this is a problem of correct classification or recognition of brain waves, which we observe at the surface of the skull by electroencephalographic (EEG) recordings of the electric field.

The concept of isomorphism of two models of a theory is one of the most basic notions used in many parts of mathematical psychology. It has been prominent in the

extensive development of the measurement theory of many different psychological phenomena. But the notion has a much wider sweep throughout mathematics and many parts of physics. The reason for this widespread application is easy to explain. An isomorphism captures the concept of two different models, or complicated sets of data, having the same structure with respect to a given set of concepts. The first point to be emphasized is that the sameness of structure is not meant to be the relation of identity for the entities in question, which is trivial, both from a mathematical and scientific standpoint. What is important is that the notion of same structure captures that part or, better, the properties or features of the objects or phenomena under consideration, that are dealt with theoretically or experimentally. In experiments, features that are not considered relevant to the analysis of same structure are, for example, the names of the experimenters or the dates of birth of the experimenters, or even the exact day on which an experiment took place. At a more fundamental level, in classical particle mechanics, to give an example from physics, the color of particles is not part of the mechanical analysis of structure, even though the concept of color is important in other parts of physics.

A relation of isomorphism must be an equivalence relation, that is, reflexive, symmetric and transitive, but for ordinary mathematical purposes, a good deal more is necessary. If we just required an equivalence relation, then any one-one mapping of one set onto another would provide an isomorphism. What is required for most structures is not only the mapping of one domain onto another, but also the preservation in the obvious sense, familiar from many examples, of the operations, relations and fixed objects that are part of the structure.

Sometimes the mapping from one domain to another is not a one-one function, but a many-one function, giving rise to what is called, in the theory of measurement, a homomorphism. Familiar examples are mappings of empirical structures into numerical ones, where distinct empirical objects or phenomena are assigned the same number. Thus, we have a homomorphism from the empirical phenomena to the numbers representing measurements.

The concern here is the existence of an isomorphism between constituents of language and the processes in the brain that are postulated to represent them. So, for example, we can start at the word level and look for the representation of each word in the brain. Already an ambiguity is present, one that is also present in the representation of words outside the brain. When we say a word, we have a process representation of that word. When we record it on a CD, we have a nonprocess representation that has the potential of giving us such a representation when we play the CD. We make this distinction in the brain as well, even if we do not have a settled view of what corresponds to the CD bit representation of words in the brain. There can even be disagreement as to how words are initially processed in the brain. One view is that words are processed electrically in the axons and synapses of neurons, whereas another is that they are processed in the electromagnetic field generated by the neurons. I will not get into this controversy here. In the experiments mentioned later, we used recordings of the electric field obtained from EEG sensors. I do emphasize that we are observing the electric field, not in any sense directly observing electric currents in individual axons or synapses.

For many kinds of concrete objects or phenomena, we must use a quantitative measure of fit between two objects—for example, in the fit between a prototype and a test sample. This leads to a notion of similarity, as a generalization of the concept of

isomorphism. This relation is, in general, not transitive. Similarity is not a new concept in psychology. It has been widely studied since the nineteenth century. Even so, it continues to be a source of formal problems. The transitivity is only the beginning of the troubles. In the ordinary concept of isomorphism and the preservation of structure, we have substitution relations for parts of a given object or phenomenon. But these straightforward substitutions, under a standard concept of congruence or isomorphism, are much more complicated when the familiar thresholds of psychological phenomena disturb the transitivity of the equivalence and the isomorphism as well.

Isomorphism of language structures. We first characterize context-free elementary language structures for the restricted purpose of formulating our hypothesis of structural isomorphism for constituents of language and their brain representations. Weaker, context-dependent structures are introduced later.

We define an *elementary segment* as a pair *(T,f)*, where *T* is a half-open, half-closed interval [*b*,*e*) and *f* is a real-valued function defined on *T*. The intended interpretation is that *T* is an interval of time, with *b* being the beginning of a phoneme, word or sentence and *e* being its temporal end. Then *f(t)* is interpreted as the amplitude at time *t* of a spoken word, for example, or, also as the amplitude of the word's brain representation for some listener who heard it.

The intuitive idea is that a word segment consists of a concatenation of phoneme segments, and a sentence, a concatenation of word segments. In the context-free version of the axioms, the congruence relation $\approx$ has the expected standard properties. For example, two spoken tokens of the word *Paris* would be congruent.

Definition 1 *A structure* $\mathcal{E} = (E, P, W, S, \mathcal{P}(P), \mathcal{P}(W), \mathcal{P}(S), \approx)$ *is an* elementary language structure *if and only if the following axioms are satisfied*:

1. *E is a nonempty finite set of elementary segments (T, f) as defined above.*
2. $E = P \cup W \cup S$.
3. $P \cap W = W \cap S = P \cap S = \mathbf{0}$.
4. *Each word w in W is a concatenation of phonemes in P, i.e.,*
   (i) $w = (T,f)$,
   (ii) $w = p_1 \cdots p_m$,
   (iii) $p_i = (T_i, f_i)$,
   (iv) $T = [b,e) = [b_1, e_1) \cup \cdots \cup [b_m,e_m)$,
   (v) $b = b_1 \cdots e_i = b_{i+1} \cdots e_m = e$,
   (vi) $f = f_1 \cup \cdots \cup f_m$.
5. *Each sentence s in S is a concatenation of words in W.*
6. *The relation* $\approx$ *is an equivalence relation on* $(P \times P) \cup (W \times W) \cup (S \times S)$.
7. *If* $w = p_1 \cdots p_m$, $w' = p'_1 \cdots p'_m$, *and* $p_i \approx p'_i$, $1 \leq i \leq m$, *then*

$$w \approx w'.$$

8. *If* $s = w_1 \cdots w_n$, $s' = w'_1 \cdots w'_n$, *and* $w_i \approx w'_i$, $1 \leq i \leq n$, *then*

$$s \approx s'.$$

9. *Any two elements of E that belong to the same member of any one of the three partitions* $\mathcal{P}(P)$, $\mathcal{P}(W)$ or $\mathcal{P}(S)$ *are congruent, independent of the context in which they occur*.

Although the axioms are stated in formal mathematical language, each one has a natural interpretation. Axiom is just a finiteness restriction on *E*, characteristic of any sample of spoken or written language, however large it may be. The infinite structures characteristic of the formal theory of grammars do not, of course, impose such a restriction. Axiom 2 guarantees that every elementary segment in *E* is either a phoneme (*p*), word (*w*) or sentence (*s*). This is sufficient for our purposes. In other contexts, it would be natural to add syllables and, for English at least, many familiar short phrases that would be better to treat as units rather than as being composed of words. Axiom 3 permits no overlap between phonemes, words and sentences. One-word sentences do not occur in the work reported here. Accommodating them in other circumstances can be handled in several different ways. Axioms 4 and 5 characterize how words are concatenated from phonemes, and sentences from words. The temporal-process construction used here is meant to be realistic for verbal stimuli that are auditory or visual, matched in display timing to the corresponding auditory stimulus; and also for the recording and analysis of their brain representations.

Axioms~4 and 5 also represent a divergence from the axiomatic formulation of formal grammars where sentences are described as finite sequences of words, and no formulation of the real-time process of production or comprehension is given. Axiom 6 is the requirement that the congruence relation $\approx$ be an equivalence relation, i.e., that it be reflexive, symmetric and transitive. Axioms 7 and 8 state the conditions that must be satisfied for a binary relation of equivalence to be a congruence relation. The congruence axioms have been written, for simplicity, in a geometric style. They can be converted to a standard algebraic form by explicitly introducing a binary operation of concatenation. Axiom 9 is conceptually important because we think of the partitions as having the property that a set which is a member of a partition can be identified as the *type* of a phoneme, word or sentence, as the case may be. Such a distinction, within a formal set of axioms for linguistic processes or objects, is unusual, but desirable here, because of the emphasis on tokens—the elementary segments—as the objects of analysis.

We now turn to the definition of isomorphism of two elementary language structures. Another concept we formalize along with isomorphism is that of an approximate isometry between corresponding elementary segments. The intuitive idea back of this constraint is that the real-time processing of the constituents of language and their brain representations must take about the same amount of time. For example, the processing of a spoken word ordinarily takes place within a few hundred milliseconds of its being heard, with a small set of troublesome words taking longer. In our experiments, all of the words used were meant to be familiar and easily processed. In a more elaborate statement of the theory, allowance needs to be made for difficult words. For example, in systematic psychological theories of eye movement during reading, if processing of a word or phrase is not complete and non-stimulus supported memory has decayed, backtracking to the

immediately preceding word or phrase is likely to occur. In so doing, a tight approximate isometry would be disturbed (for detailed axioms on eye movements in reading, see Suppes, 1990).

Definition 2 Let $\mathcal{E}_1 = (E_1, P_1, W_1, S_1, \mathcal{P}(P_1), \mathcal{P}(W_1), \mathcal{P}(S_1), \approx_1)$ *and* $\mathcal{E}_2 = (E_2, P_2, W_2, S_2, \mathcal{P}(P_2), \mathcal{P}(W_2), \mathcal{P}(S_2), \approx_2)$ *be two elementary language structures.* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *are* isomorphic i*f and only if there exists a one-one mapping* φ *from* E_1 *onto* E_2 *such that*

(i) $\varphi(P_1) = P_2$, $\varphi(W_1) = W_2$, $\varphi(S_1) = S_2$ *and the partitions of* P_1, W_1 *and* S_1 *are also preserved under the mapping* φ;

(ii) *For* e *and* e' *in* E_1,

$$\varphi(e) \approx_2 \varphi(e') \text{ iff } e \approx_1 e';$$

(iii) *If* $w = p_1 \cdots p_m$, *then* $\varphi(w) = \varphi(p_1) \cdots \varphi(p_m)$;

(iv) *If* $s = w_1 \cdots w_n$, *then* $\varphi(s) = \varphi(w_1) \cdots \varphi(w_n)$.

Moreover, the mapping φ *is an* ε*-approximate isometry if for any element* (T,f) *of* E_1 *with* $T = [b,e)$

$$|(e - b) - (\varphi(e) - \varphi(b))| < \varepsilon.$$

In some occurrences of spoken words and sentences, the context matters. A phoneme changes in the context of different phonemes, and similarly for words. To accommodate and study these phenomena, we weaken the axioms for context-free structures to context-dependent ones. But these modifications will not be considered here.

So far in this section I have not said how to characterize congruence for elementary segments, i.e., phonemes, words and sentences, of brain representations of language. Keep in mind that this characterization is critical for data analysis, and does always refer to tokens, not types, of linguistic units, contrary to abstract versions of phonology and syntax. Some form of least-squares fit, summed over the discrete amplitude observations of the two electromagnetic waves, is mainly used in the models we tested. This means that congruence between test samples and prototypes is judged by which prototype fits the test sample best.

This is not the occasion to enter into the experiments in which data were collected and analyzed to test the hypothesis of structural isomorphism of language constituents and the brain representations of them, i.e., phonemes, words, and sentences in the present context. I can display some graphic data. Figure 1 shows the close similarity of the brain representations of the word *east* based on data from different sentences. The close similarity supports, of course, the existence of a structural isomorphism, even though these data make nothing like a complete case. Figure 2 shows the same comparison of brain representations for the four consonants *b, g, p,* and *t,* when they occurred as initial consonants. To avoid any misunderstanding, I stress that structural isomorphism between language constituents and their brain representations does not mean that a brain

representation of a spoken word, for example, perceptually resembles it. The most vivid example to make this same point is the great physical and perceptual distance between actual empirical procedures of measurement and their isomorphic representation by numerical structures.

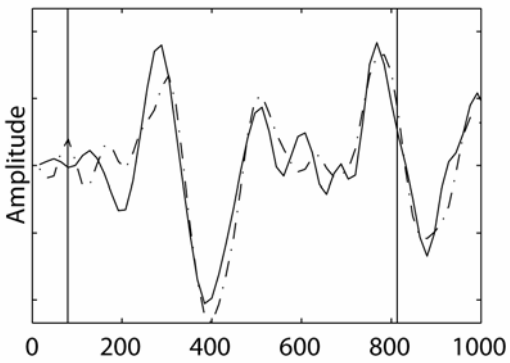

Figure 1. The prototype for *east* is the solid curved line, and the dashed-and-dotted line the test sample from an experiment. The *x* axis is measured in milliseconds from the onset of the visual word as stimulus. The *y* axis is measured in microvolts, with the numerical scale not shown. The vertical line on the left shows the beginning *b* and the one on the right the end *e*, thus marking off the temporal segment used for the least-squares test of fit.

If the hypothesis of structural isomorphism were blatantly wrong, it would seem to make the brain's methods of computation in recognizing external processes and objects much more complex than they probably are. Structural-isomorphism claims are restricted, but still supportive of the general idea that the brain representations of processes and objects are similar, in this strong structural sense, to what they represent. Detailed empirical findings on such an isomorphism are given in *Several models to test the hypothesis of structural isomorphism between constituents of language and their brain representations* (Suppes et al., in preparation).

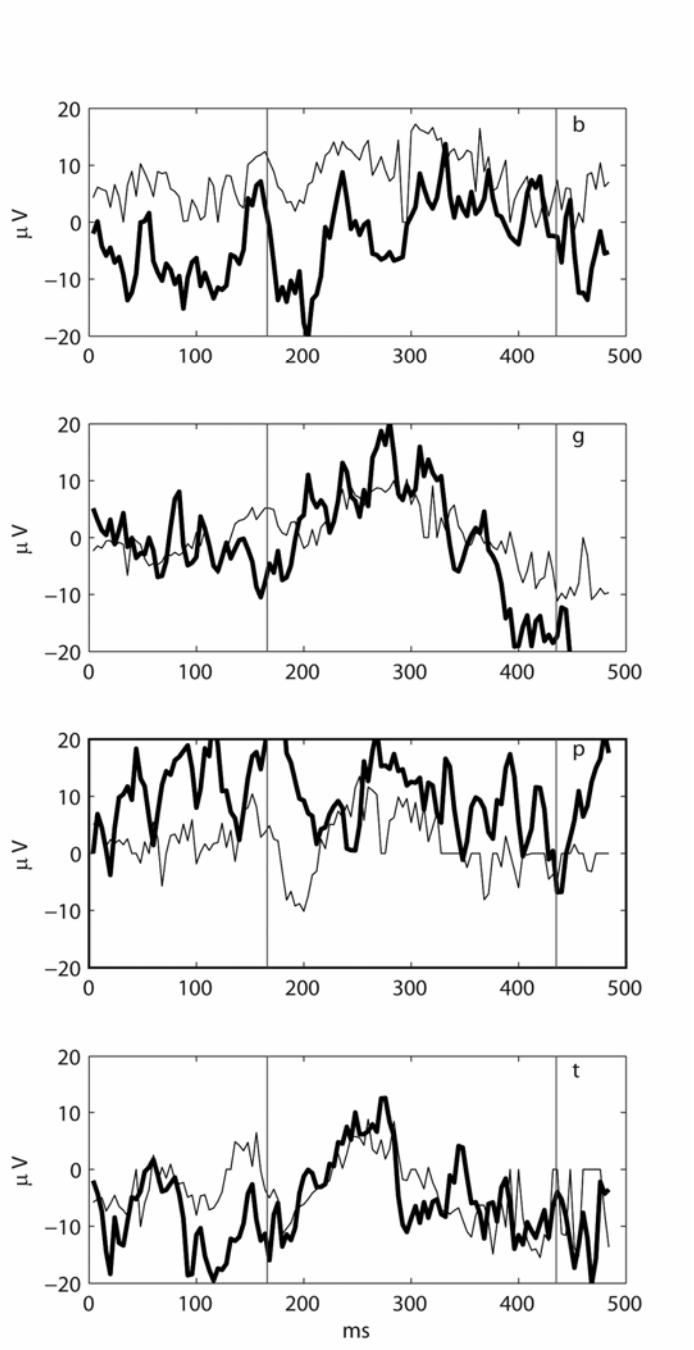

Figure 2. Comparison of prototype and typical test sample after data censoring for each consonant. The thick solid line is the prototype and the thin solid line a typical test sample, both after censoring. The measurements on the *x* and *y* axes are the same as for Figure 1.

III. Habits, Automaticity and Consciousness

Our mental concept of ourselves is that of self-aware human beings able to think deliberately, carefully, and thoroughly about all kinds of problems. But contrary to this folklore ideal, and contrary also to much folklore psychology, we are almost entirely unaware or unconscious of our thinking processes. What we often have is excellent knowledge of the results. We know, for example, that we have decided to go to the movies, or we have now made the decision to take a trip to the Arctic the next time we see a special fare. These results are salient, often privately known only to the individual having them, but all the same they reflect the results, not the patterns of thinking that led to them. It is easy to say, "Well, this is just another philosophical opinion." In fact, the data are quite substantial in supporting the empirical conclusion that we have little consciousness of process, but much of results. Two useful articles providing many references to supporting empirical studies are Nisbett and Wilson (1977) and Wilson (1985). I have also surveyed these studies myself in an earlier article, Suppes (2003).

Most of our walking and talking proceeds a pace without any need for prior deliberation or conscious reflection. If we stumble, we notice it, and are momentarily conscious of the steps we are taking. If we mumble when we talk, we are in like manner suddenly aware that we have made a mistake, and need to correct it. But these are exceptions. The normal course of things is wonderfully unselfconscious and smoothly running, without any awkward interference of efforts at deliberate and conscious thought.

But clearly, there must be a place for deliberation and conscious reflection. Of course there is. Partly, it is in the contemplations, happily or sadly, of the results of these unconscious processes. Another, and more important, occasion for such consciousness is in the learning of something new and the deliberate effort to pay attention, to concentrate, to control movements, thoughts, and so forth. Such initial learning of habits is in many ways the most significant case of conscious attention.

From Aristotle to James. The ideas that I am expressing about habits and about consciousness are not presumed to be new. In an excellent recent article Burnyeat (1999) points out how fundamental for Aristotle is the importance of habit in learning how to be good. Those who have stressed the great importance of the practical syllogism for Aristotle's analysis of the good have too often missed this point. Only the properly prepared person, i.e., someone raised from early youth in the proper manner, is going to be able to achieve the desired state of flourishing described so well by Aristotle. Burnyeat's emphasis on the importance of the psychological development of the young as a central and essential feature of Aristotle's theory of virtue is a welcome and fresh addition to many of the standard readings of the *Nichomachean Ethics*.

Equally important is Burnyeat's emphasis that, in Aristotle's view, learning what is noble and just, that is, learning virtue, does not consist simply of learning neat formulations of rules or traditional maxims. It takes an educated perception, the habitual capacity of going beyond the application of general rules, to know what is needed for the practice of the virtues in specific circumstances (*Nichomachean Ethics*, 1109b23, 1126b2-4). A similar thesis about Aristotle is developed by Sherman (1999), who also quotes relevant passages from the *Politics*, *De Anima*, *Poetics*, and *Rhetoric*, to support the importance of the developmental role of learning good habits early.

That the Greek teaching of the young by tutors or in elementary schools followed the precepts of Aristotle is, of course, only partly true. A rather realistic account of Greek education in Hellenistic and Roman times may be found in Cribiore (2001). The one thing that is reinforced by her complex account of what the practice was like is the emphasis on the formation of habits. To a very large extent, the teaching in the early years emphasized mental gymnastics, as an analogue of the recognition of the need for physical gymnastics in the training of the body.

Much more detailed than Aristotle's own treatment of habits is that of Aquinas. I shall not try to survey it here. Some attention to it is given in Drolet and Suppes (in press). Hume does not much use the word "habit", but writes about custom and here he is close to the earlier tradition. His definition is very much in the spirit of what I have been saying here about habits. "We call everything custom which proceeds from a past repetition without any new reasoning or conclusion" (*Treatise*, p. 104). In passages on the same page and others close by, Hume very much reinforces the kind of automaticity of habits that I want to stress more explicitly a little later. He makes the important point that in very obvious uniform conjunctions of events, what we then infer are causes and their effects; the mind does not reflect at all or reason upon the phenomena. It passes from one to the other, driven by the constant association. Here is the passage that is clear on this point:

> In general we may observe, that in all the most establish'd and uniform conjunctions of causes and effects, such as those of gravity, impulse, solidity, &c., the mind never carries its view expressly to consider any past experience: Tho' in other associations of objects, which are more rare and unusual, it may assist the custom and transition of ideas by this reflexion.
>
> (Hume, *Treatise*, p. 104)

But the *locus classicus* on this topic is Chapter 4 of James' *Principles of Psychology* (1890). Here is the important passage about the move from conscious attention to automaticity, James point being that "habit diminishes the conscious attention with which our acts are performed". This is what he has to say about this process:

> One may state this abstractly thus: If an act require for its execution a chain, *A, B, C, D, E, F, G,* etc., of successive nervous events, then in the first performances of the action the conscious will must choose each of these events from a number of wrong alternatives that tend to present themselves; but habit soon brings it about that each event calls up its own appropriate successor without any alternative offering itself, and without any reference to the conscious will, until at last the whole chain, *A, B, C, D, E, F, G,* rattles itself off as soon as *A* occurs, just as if *A* and the rest of the chain were fused into a continuous stream. When we are learning to walk, to ride, to swim, skate, fence, write, play, or sing, we interrupt ourselves at every step by unnecessary movements and false notes. When we are proficients, on the contrary, the results not only follow with the very minimum of muscular action requisite to bring them forth, they also

> follow from a single instantaneous ‘cue’. The marksman sees the bird, and, before he knows it, he has aimed and shot. A gleam in his adversary's eye, a momentary pressure his rapier, and the fencer finds that he has instantly made the right parry and return. A glance at the musical hieroglyphics, and the pianist's fingers have rippled through a cataract of notes. And not only is it the right thing at the right time that we thus involuntarily do, but the wrong thing also, if it be an habitual thing. Who is there that has never wound up his watch on taking off his waistcoat in the daytime, or taken his latchkey out on arriving at the door-step of a friend? Very absent-minded persons in going to their bedroom to dress for dinner have been known to take off one garment after another and finally to get into bed, merely because that was the habitual issue of the first few movements when performed at a later hour. The writer well remembers how, on revisiting Paris after ten years’ absence, and, finding himself in the street in which for one winter he had attended school, he lost himself in a brown study, from which he was awakened by finding himself upon the stairs which led to the apartment in a house many streets away in which he had lived during that earlier time, and to which his steps from the school had then habitually led.
>
> (James, *Principles of Psychology*, p. 114–115)}

We do not have to be explicit about the conscious will, as James is in this passage, to appreciate fully the empirical correctness of what he says about forming habits. He does not use the term ‘automaticity’ that is now popular in the psychological literature, but it is clear that he has in mind exactly this phenomenon. A habit becomes automatic when it is performed without conscious attention. So in that sense consciousness and automaticity are opposites. Well, so what? The answer is what James stresses and other psychologists of the present time stress. Automaticity is the road to perfection. The conscious performance of acts that we think of as habitual is awkward, badly timed, and often inappropriate. The point of this remark is to emphasize how much daily experience, the amount of which is not fully appreciated by almost any philosophers of mind, is based on automatic, unconscious and habitual responses. Not on conscious deliberations or reflections of any kind. Moreover, this is true of our most cognitive activities, talking and listening. We automatically process the speech we produce and the speech we listen to. In almost no cases are we really conscious of either process. It is only after we have turned over in our minds what we have said, or what we have heard, that some point will come sharply to conscious attention. Again, this is a matter of process being mainly unconscious, but partial results being available for conscious inspection.

Very much in the Aristotelian spirit of developing good habits is the following passage from James urging exactly that at a young age:

> The great thing, then, in all education, is to *make our nervous system our ally instead of our enemy*. It is to fund and capitalize our acquisitions, and live at ease upon the interest of the fund. *For this we must make automatic and habitual, as early as possible, as many useful actions as we can*, and guard against the growing into ways that are likely to be disadvantageous to us, as we should guard against the plague. The more of the details of

> our daily life we can hand over to them effortless custody of automatism, the more our higher powers of mind will be set free for their own proper work. There is no more miserable human being than one in whom nothing is habitual but indecision, and for whom the lighting of every cigar, the drinking of every cup, the time of rising and going to bed, every day, and the beginning of every bit of work, are subjects of express volitional deliberation. Full half the time of such a man goes to deciding, or regretting, of matters which ought to be so ingrained in him as practically not to exist for his consciousness at all. If there be such daily duties not yet ingrained in any one of my readers, let him begin this very hour to set the matter right.
>
> (James, *Principles of Psychology*, p. 122)

James goes on to strengthen this passage by emphasizing that in deciding to launch ourselves into a new habit, we must be as decisive and determined as possible. So, the place for consciousness in the initial change of habits is a salient and important one. But salience and importance recede, as should be the case, once the habit becomes automatic. What is good about this discussion of the learning of habits on James' part is his full recognition that habits are not something that is purely mental in some dualistic Cartesian sense, but always intimately related to both past and present actual performance. This is what James says:

> Down among his nerve-cells and fibres the molecules are counting it, registering and storing it up to be used against him when the next temptation comes. Nothing we ever do is, in strict scientific literalness, wiped out. Of course, this has its good side as well as its bad one. As we become permanent drunkards by so many separate drinks, so we become saints in the moral, and authorities and experts in the practical and scientific spheres, by so many separate acts and hours of work.
>
> (James, *Principles of Psychology*, p. 127)

What he also fully recognizes is that the basic physiological act of learning habits has remarkable similarity whether habits are good or bad. In other respects, as is emphasized in Drolet and Suppes (in press), the differences between good and bad habits are well accepted and much agreed upon in broad communities of persons ranging from the young to the old.

Consciousness. In the later part of the chapter on habits, James examines several ways of thinking about the function of consciousness. I will not try to go through his various arguments. He does have, however, at the end a very suggestive idea. It is a functional one, not one identified by any particular physical or neural manifestation of consciousness. The idea is that the role of consciousness is a selective one. When we are confronted with a number of alternatives it is the function of consciousness to help us think about the choice we will make. Notice the difference here. When automaticity governs, the choice is made smoothly and unconsciously. When the choice is not one governed by habit, then consciousness often, even if not always, comes in to play.

It is useful to compare James' notion of selection with the notion that was dominant a few years later, namely, attention. By the first decade of the twentieth century it had become a central concept of experimental psychology. Some standard references are Wundt (1897), Pillsbury, (1908), and Titchener (1908). One leading idea of this functionalist approach to attention was that the initial condition for onset is stimulus change, and the recognition that the evidence of attention was also manifested in various physiological responses. However, the tight connection of attention to consciousness was subsequently lost in the pursuit of a strict behaviorism, especially in American psychology. It was regarded as unscientific in most of the decades of the first half of the twentieth century for experimental psychologists to speak of consciousness.

But the study of attention prospered not only in the United States but also in Europe, especially in Russia, then the Soviet Union, for example, in the work of Luria and Vinogradova (1959) and Sokolov (1960, 1963). The Russian psychologists tended to talk about the orienting reflex rather than attention, but the functional conception was very similar. Such studies were also made in the United States. A good example is the study of the relation of the orienting reflex conditioning in Maltzman and Raskin (1965). Two good, more recent references on automaticity are Schriffin and Schneider (1977) and Barge and Chartrand (1999).

In spite of the new neural and psychological studies of consciousness, many questions do not yet have good answers. What remains constant is the psychological insistence that phenomenological awareness is the central feature of consciousness. This is true wherever awareness occurs and whether it is viewed as essential or accidental on any particular occasion of feeling, perceiving or thinking. Although not without controversy, the experimentally supported stress on awareness of results, not processes, will come to be recognized as sound, and will help narrow the phenomenological range of consciousness. I am skeptical that the physical source of consciousness will be found in any one location in the brain. Much more promising in my view is the conjecture that consciousness reflects a physical process of activation, as briefly described in Part I, and it may occur in many places in the brain, at least in the cortex. For an extensive review of the experimental literature on the relation between activation and consciousness, see Dehaene and Changeux (2005). They also support the hypothesis of a conscious neuronal workspace that distinguishes automatic subsystems from more autonomous, spontaneous supervisory systems. Their ideas build on the earlier cognitive theory of consciousness of Baars (1989). These ideas are very suggestive but still speculative. My own, also speculative, idea is that the physical embodiment of consciousness will be more dynamic and electromagnetic, with exact spatial location being of less importance.

References

Aquinas, T. (1944). Summa theologica. In A. C. Regis, Ed., *Basic writings of Saint Thomas Aquinas*, Vol. 1. New York: Random House.

Aristotle (1941). Metaphysics. In R. McKeon, Ed., *Basic works of Aristotle*. New York: Random House.

Aristotle (1941). Nichomachean ethics. In R. McKeon, Ed., *Basic works of Aristotle*. New York: Random House, pp. 1109b23, 1126b2–4.

Aristotle (1975). *De Anima (On the soul)*. Cambridge, MA: Harvard University Press, 4th edn. English translation by W. S. Hett. First Published 1936.
Baars, B. J. (1989) *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.
Barge, J. and T. Chartrand (1999). The unbearable automaticity of being. *American Psychologist*, **54**: 462--479.
Brennan, R. E. (1941). *Thomistic psychology: A philosophical analysis of the nature of man*. New York, NY: Macmillan.
Brown, C. M. and Hagoort, P. (1999). *The neurocognition of language*. Oxford: Oxford University Press.
Burnyeat, M. F. (1999). Aristotle on Learning to be good. In N. Sherman, Ed., *Aristotle's ethics*. New York: Rowman and Littlefield.
Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.
Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, **11**, 453–482.
Cribiore, R. (2001). *Gymnastics of the mind, Greek education in Hellenistic and Roman Egypt*. Princeton: Princeton University Press.
Dehaene, S. and J. P. Changeux (2005). Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. *PLoS Biology*, **3**: 0910–0927.
Descartes, R. (1649/1927). *Passions of the soul*. In R. M. Eaton, Ed., *Descartes selections*, pages 361-403. New York, NY: Charles Scribner's sons. First published 1649.
Drolet, A. and P. Suppes (in press). The good and the bad, the true and the false. In M. C. Galavotti, R. Scazzieri, and P. Suppes, Eds., *Reasoning, rationality, and probability*. Stanford, CA: CSLI Publications.
Hume, D. (1739/1951). *A treatise of human nature*. London: John Noon. Quotations taken from L. A. Selby-Bigge's edition, Oxford University Press, London.
James, W. (1890/1918). *The principles of psychology*. New York: Henry Holt and Company.
Lear, J. (1978). Going native. *Daedalus*, **107**, 177–78.
Luria, A. R. and Olga S. Vinogradova (1959). An objective investigation of the dynamics of semantic systems. *British Journal of Psychology*, **50**, 89–105.
Maltzman, I. and D. C. Raskin (1965). Effects of individual differences in the orienting reflex on conditioning and complex processes. *Journal of Experimental Research in Personality*, **1**, 1–16.
Minsky, M. L. (1967) *Computation: finite and infinite machines*. Englewood Cliffs, New Jersey: Prentice Hall.
Nisbett, R. E., and T. D. Wilson, (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, **84**, 231–259.
Pillsbury, W. B. (1908). *Attention*. New York: The Macmillan Company.
Rugg, M. D., and Coles, M. G. H. (Eds.). (1995). *Electrophysiology of mind: Event-related brain potentials and cognition*. Oxford: Oxford University Press.
Schriffin, R. M. and W. Schneider (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, **84**, 127–190.

Sherman, N. (1999). The fabric of character. In N. Sherman, Ed., *Aristotle's ethics*. New York: Rowman and Littlefield.

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. New York: The Macmillan Company.

Sokolov, E. N. (1960). Neuronal models and the orienting reflex. In Mary A. B. Brazier, Ed., *The central nervous system and behavior*. New York: Josiah Macy, Jr. Foundation, pp. 187–276.

Suppes, P. (1974). Aristotle's concept of matter and its relation to modern concepts of matter. *Synthese* **28**, 27–50.

Suppes, P. (1990). Eye-movement models for arithmetic and reading performance. In E. Kowler (Ed.), *Reviews of oculomotor research*, (Vol. IV), *Eye movements and their role in visual and cognitive processes* (pp. 455–477). New York: Elsevier.

Suppes, P. (2002). *Representation and invariance of scientific structures*. Stanford, CA: CSLI Publications.

Suppes, P. (2003). Rationality, habits and freedom. In N. Dimitri, M. Basili, and I. Gilboa, Eds., *Cognitive processes and economic behavior*. Routledge Siena Studies in Political Economy. New York: Routledge, pp. 137–167.

Suppes, P. and J-Y Béziau (2004). Semantic computations of truth, based on associations already learned. *Journal of Applied Logic*, **2**, 457–467.

Suppes, P., M. P. Guimaraes, D. K. Wong, and E. T. Uy (in preparation). Several models to test the hypothesis of structural isomorphism between constituents of language and their brain representations.

Suppes, P., B. Han, J. Epelboim, and Z.-L. Lu (1999a). Invariance between subjects of brain wave representations of language. *Proceedings National Academy of Sciences USA*, **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 National Academy of Sciences USA*, **96**, 14658–14663.

Suppes, P., Z.-L. Lu, and B. Han (1997). Brain wave recognition of words. *Proceedings National Academy of Sciences USA*, **94**, 14965–14969.

Tarski, A. (1936). The concept of truth in formalized languages. In A. Tarski, *Logic, semantics, metamathematics*. Translated by J. H. Woodger. Second edition, 1983, edited by John Corcoran. Indiana: Hackett Publishing Company, Inc.

Themistius (350 A.D./1996). *On Aristotle's On the Soul*. Translated by Robert B. Todd. New York: Cornell University Press.

Titchener, E. B. (1908.) *Lectures on the elementary psychology of feeling and attention*. New York: The Macmillan Company.

Wilson, T. D. (1985). Strangers to ourselves: The origins and accuracy of beliefs about one's own mental states. In J. H. Harvey and G. Weary, Eds., *Attribution: Basic issues and applications*. Orlando, FL: Academic Press.

Wundt, W. (1897/1902). *Outlines of psychology*. Leipzig, W. Engelmann; New York, G. E. Stechert, 2nd (revised) English edn., from the 4th (revised) German edn., translated by C. H. Judd.