Principles for Implicit Learning

Axel Cleeremans

Séminaire de Recherche en Sciences Cognitives
Université Libre de Bruxelles CP 122
Avenue F.-D. Roosevelt, 50
1050 Bruxelles — BELGIUM
axcleer@ulb.ac.be

April, 1996

Cleeremans, A. (1997). Principles for Implicit Learning. In D. Berry (Ed.), How implicit is implicit learning? (pp. 196-234), Oxford: Oxford University Press

The author is a Research Associate of the National Fund for Scientific Research (Belgium). I thank Mark St.John and Pierre Perruchet, who provided valuable reviews on an earlier version of this paper. I also thank Alain Content, Bob French, Luis Jiménez and Tim Van Gelder, all of whom contributed critical and stimulating discussion about the issues. Portions of this chapter were adapted from Cleeremans (1994) and from Cleeremans (1995).

1. Introduction

Implicit learning research has now reached a point of unprecedented interest. After about 25 years, this field—which once appeared as if it would stay relatively marginal for ever—is currently witnessing a tremendous revival, with a steady outflow of new studies, many new authors contributing actively, and many new bridges to related fields. Yet, despite all the current excitement and the significant advances the field has made recently, we seem to be only somewhat closer to providing answers to the questions that Reber started asking himself around 1967 (e.g., Reber, 1967). Indeed, the field still appears to be divided over central issues, such as whether implicit learning entails the possibility of unconscious cognition, or whether knowledge acquired implicitly possesses properties (e.g., its potentially abstract character) similar to those that characterize knowledge acquired explicitly.

Consider for instance Reber's claim (Reber, 1993) that in order to demonstrate tacit knowledge for any stimulus domain, it is sufficient to show that a>b, where a is the sum of information available to the unconscious and b is the sum of information available for conscious expression. Many authors take issue with this statement (e.g., Shanks & St.John, 1994), and argue instead that in order to demonstrate implicit knowledge, one also needs to establish that a>0 and that b=0, that is, not only that behavior is influenced by unconscious determinants, but also, and most importantly, that conscious knowledge of the relevant information is nil. Other authors strongly disagree with this position, claiming essentially that it is impossible to demonstrate that b=0, that attempts to do so are futile, and that this position is thus untenable.

Even though it appears technically obvious that implicit learning has not been established as long as b>0, I find myself very much in agreement with Reber's position on this issue, for several reasons. First, it seems clear that any theory of cognition has to make room for a concept such as "implicit". It seems utterly implausible to assume that all we learn or process is consciously available or intentional. Second, and perhaps paradoxically at first, associations rather than dissociations between implicit and explicit knowledge are expected with normal subjects. The numerous recent findings (see Shanks & St.John, 1994) that participants are in fact aware of some information that was previously assumed to be implicit do not necessarily shatter the notion that learning can be implicit. They merely indicate that when explicit knowledge is assessed in a more sensitive way, it turns out that participants can express it. This, however, has no bearing on whether the information thus revealed is acquired or even used explicitly. It may therefore turn out to be impossible to show that b=0 not only because of methodological problems, but also simply because b is never equal to zero in normal subjects: Concscious awareness cannot simply be turned off.

Why then would defenders of implicit learning insist on having two learning systems when most of the evidence suggests that in normal subjects implicit and explicit learning are associated? Would it not be simpler to merely assume that behavior is always determined by conscious contents? This is a theoretical position that has found many supporters recently, yet it appears to be ultimately incomplete to many others. Reber (1993) aptly summarizes this feeling by stating that the point of implicit learning research is not "...to show that consciousness is totally absent from the process but instead that explanations of behavior based solely on conscious factors do not provide satisfactory accounts." (p. 70). Reber (1993) recommends that we adopt what he calls "the implicit stance", that is, the belief that unconscious cognition is the default mode of processing, and that the burden of proof should be put on supporters of "conscious cognition" rather than on "implicit-learning-as-default" people. However much I

agree with Reber that implicit cognition is the default, I believe that taking it as axiomatic may be a counter-productive strategy. Instead, I believe that one can identify principles that make it less problematic for implicit cognition to be the default. For instance, one such principle may be that the implicit/explicit distinction does not necessarily reflect an architectural dichotomy. The arguments described in the previous paragraphs may thus be flawed because they require the assumption that a given piece of knowledge is either in the "unconscious" box or in the "conscious" box. However, there may simply be no such boxes.

More generally then, I want to argue that the controversies that divide the field today find their roots in models of cognition that fail to capture the complexity of the phenomena they are meant to explain. In short, I think that much of the current debate comes about because we still tend to think about cognition as emerging through the operations of a symbolic processor that essentially fetches information from separable knowledge databases, processes it, and then sends it to some other module for further processing or action. This "warehouse, truck, and factory" metaphor of cognition (McClelland, personal communication, 1991) is in need of revision, and the goal of this chapter is to explore alternative ways to think about some of the central issues in implicit learning research.

To do so, I start by sketching the "classical metaphor of cognition", which is rooted in modularity and symbol processing. I then proceed to show how this framework leaves no room for the concept of implicit, short of (1) ascribing implicit learning to a separate "implicit learning system" — a theory that I call the "shadow" theory of implicit learning, or (2) denying implicit learning altogether. I show that both attempts to reconcile the empirical data with the classical framework are unsatisfactory, and suggest that the framework itself is flawed because the framework, and the research practices inspired by its tacit adoption, rely heavily on a cluster of assumptions about the relationship between behavior and the underlying cognitive system which I collectively refer to as the "assumptions of direct mapping". These assumptions are basically variations on the theme that there is a direct and transparent relationship between observable patterns of behavior and the internal representations and processes that produce them.

I then proceed to introduce four overlapping ways in which these assumptions can be violated and illustrate each with data and arguments taken from implicit learning research and from computational modeling. Each of these four possible violations of direct mapping instantiates a corresponding principle, as follows: First, sensitivity to some regularity does not necessarily entail that this regularity is represented as a manipulable object of representation by the cognitive system. For instance, observing that recall in a memory task is organized in chunks does not necessarily entail that representations are chunked also. Second, modularity may only be functional. This means that observing a double dissociation between two behaviors does not necessarily entail that there are separable modules underlying each behavior. Third, tasks are not process-pure. Contrary to a widely held but, on the face of it, rather implausible assumption, it appears illusory to assume that one can devise measures of some performance that involve only one processing module. Fourth, many dimensions of cognition that are often cast as static and dichotomous may in fact often turn out to be dynamic and graded.

These issues may appear to be moot points about implicit learning in that they seem to leave room for every kind of theoretical position, but when embodied together within a computational theory, they turn out to be powerful principles with which to think about human cognition in general, and implicit learning in particular.

The goal of this chapter, then, is not so much to provide a critical assessment of current methodology and thinking in the implicit learning field (see for instance, Perruchet & Gallego, this volume, for a much more complete treatment of the empirical issues), but rather to summarize and organize what I believe some of the most difficult issues may be, and to provide an alternative framework to think about these issues. A side effect of this endeavour is that it turns out, perhaps unsurprisingly, that many of these issues appear to be better addressable in the connectionist framework than in other theoretical frameworks. Hence I also take this opportunity to expose again some of the principles that I think mandate connectionism as the framework of choice for thinking about and for understanding the mechanisms involved in implicit learning.

2. The classical metaphor of cognition

The central argument of this chapter is that implicit learning is problematic only because the traditional framework within which we think about cognition is flawed. This "classical metaphor of cognition" (e.g., Newell & Simon, 1972; Fodor, 1983; Fodor & Pylyshyn, 1988) goes roughly like this: There is a central processor that fetches or stores information in knowledge bases, and processes it. The processor interacts with the world through input/output systems. These subsystems are modular, that is, autonomous and informationally encapsulated. Knowledge (either "programs" or "data") is represented symbolically.

What is the problem with this characterization of cognition, and is there an alternative? In an interesting paper, Bates and Elman (1992) highlighted several important differences between the classical, symbolic framework described in the previous paragraph and the connectionist framework, which they respectively describe as the first and second computer metaphors of cognition. It is worth going over their analysis in some detail here.

First, representations in the traditional model are discrete. As Bates and Elman (1992) put it, in a symbolic system, "...there is no such thing as 50% of the letter A or 99% of the number 7" (p. 6). Human cognition, in stark contrast, is obviously characterized by highly flexible representations that allow us to process partial or degraded information, such as blotted characters or strange foreign accents, with remarkable ease. Discrete representations do not per se prevent processing of such information, but they complicate considerably the mechanisms required to give flexibility to the resulting systems.

Second, rules in traditional models tend to be absolute: A given rule will either apply or fail to apply to the current situation. Granted, contemporary symbolic models use many ways of overcoming this limitation, for instance by assigning weights to rules or by allowing fuzzy rather than absolute matches, but as Bates and Elman (1992) stress, these features are not natural properties of the architecture and often have to be tuned externally. Again, human cognition appears to be considerably more flexible in this respect.

Third, learning is viewed essentially as programming or as memorizing. When a production system acquires new knowledge, it is not because the system is self-organizing, but because the new knowledge is the product of a process of hypothesis-testing. Because the space of potential hypotheses is necessarily constrained by the architecture's original knowledge, this approach almost makes nativism mandatory. In other words, classical models do not appear to be very useful in helping us understand processes of change. Connectionist models, on the other hand, learn continuously through experience.

Finally, the first computer metaphor explicitly introduces a distinction between hardware and software, an approach that ultimately results in adopting a purely functionalist stance on cognition. By contrast, one of the central characteristics of connectionist models is that the structures that support processing are identical with the structures that support representation: The machine and what it knows are one and the same.

I should point out that some of the properties spelled out by Bates and Elman only apply to the most traditional symbolic systems, perhaps best exemplified by early examples of production systems (see for instance, Newell & Simon, 1972; Newell, 1980; Anderson, 1983). Recently, considerably improved frameworks have appeared (see Newell, 1990) that make weaker assumptions about the nature of processing and representation. By the same token, it should also be clear that many important aspects of human cognition are well captured by symbolic frameworks: Symbol manipulation, as instantiated in problem-solving and perhaps some aspects of language processing, is obviously crucial in understanding human behavior. The question is: Does assuming that cognition is about symbol manipulation help us understand basic cognitive processes as well? Taking implicit learning as an instance of such basic cognitive processes, what are the consequences of embracing the "traditional" framework for thinking about cognition? In the following section, I describe what I take to be one of the most problematic aspects of the symbolic metaphor: It leaves no room for the implicit.

3. The classical framework leaves no room for the implicit

The main argument of this section is that symbol systems can not represent implicit knowledge1. Before developing this argument, I need to define what I mean by implicit knowledge. This of course is a difficult task, but for now I will adopt the following working definition:

> "At a given time, knowledge is implicit when it can influence processing without possessing in and of itself the properties that would enable it to be an object of representation. Implicit learning is the process by which we acquire such knowledge."

I need to clarify two issues about this definition before moving on. The first is the qualification "at a given time". I believe, along with Searle (1992), that all the knowledge we possess is at least potentially accessible to consciousness, or else that this knowledge is not mental. Hence the only way for knowledge to be implicit is for it to be implicit at some particular time, that is, with respect to some specific context.

The second issue is what I mean by "object of representation". Addressing this issue in any detail is far beyond the scope of this paper, but what I want the capture by using this expression is the difference between, for instance, a thermostat that is hard wired to turn on the furnace when the temperature drops below a set point, and the same thermostat as implemented in the form of a computer program, say, a production system. If both instances of thermostat devices are functionally equivalent in that they perform the same task, it is clear that only the computer-program thermostat can be said to have a representation of the temperature setting process, which may for instance be implemented as a symbolic rule. This rule is an object of representation in the sense that it exists independently from the hardware and because it could easily be manipulated by the computer-program thermostat independently and for purposes

other than setting the temperature, something that the hard-wired thermostat remains in and of itself incapable of doing (Note that this does not entail that implicit representations should be considered as non-mental. I return to this point in section 4.1.)

My claim is then that symbolic systems can not represent implicit knowledge because the representations that the system possesses of some material always have the property that they can be objects of representation themselves. This property of symbolic representations comes about because of two distinct central assumptions that characterize symbol systems: The fact that the structures that support processing are distinct from the structures that support representation, on the one hand, and the fact that representations are compositional, on the other hand.

To start with the first point, symbol systems typically consist of a processor that interprets symbolic expressions, either programs (e.g., production rules) or data. These representations are stored in the system's memory, for instance in the form of a list of rules, or in the form of an associative network of symbolic expressions. The fact that the processor is distinct from the representations that it manipulates automatically entails that these representations can themselves be objects of representation. Symbols require the notion of distance put forward by Newell (1990), that is, a symbol is a mechanism to obtain distal access to some knowledge. In Dienes & Perner's words (Dienes & Perner, in press), such representations appear to have at least the potential to be "mental-state explicit", because the system that uses them could always decide whether or not it possesses them.

To see this, consider a symbolic model such as Ling and Marinov's symbolic model of sequence processing (Ling & Marinov, 1994), which they applied to modeling data from Lewicki, Czyzewska and Hoffman (1987). Participants in Lewicki et al.'s study were exposed to a matrix-scanning trial, which, for the purposes of this discussion, can be thought of as a variation on simpler sequence learning paradigms in which participants are asked to react to the appearance of successive stimuli as fast as possible by pressing on the corresponding key. The typical finding is that reaction times are faster for stimuli that are predictable in the context set by previous elements of the sequence than for random stimuli. Ling and Marinov's account of performance in such tasks consists of assuming that participants progressively learn a symbolic representation of the contingencies present in the stimulus material in the form of a decision tree that associates each sequence element with its successors and that can readily be translated into simple production rules such as "IF element v1 = 1 and element v3 = 1 and element v4 = 4 THEN element c = 1".

The point I want to make is as follows: Because these expressions are static and exist independently of the processor that interprets them, they are automatically available to outside inspection. Indeed, it is almost a defining feature of symbols that they require external mechanisms to interpret them. Because of this, it would thus be trivial to augment Ling and Marinov's model with procedures that enable it to justify each decision it makes, for instance. Representing knowledge with symbolic expressions thus appears to necessarily entail the possibility of accessing them for purposes other than the purpose for which they were acquired or developed, simply because they already require external mechanisms for the system to use them at all.

For a system to have the capacity to analyze its representations in this way, however, these representations also need to exhibit a second property: Compositionality, in a specifically

concatenative way. According to van Gelder (1990), an item is "said to have a compositional structure when it is built up, in a systematic way, out of regular parts drawn from a certain determinate set; those parts are then the components or the constituents of the item" (p. 356). As van Gelder (1990) points out, there are several different ways in which one can combine constituents so that the resulting complex representations are compositional. In typical symbolic systems, the way elementary constituents are combined to produce complex, compositional representations that have an internal structure is by concatenation, that is, by juxtaposition. This characteristic of symbolic systems turns out to be crucial for the arguments that I am developing here because, as van Gelder describes, concatenation, by definition, preserves the tokens of a complex expression's constituents and their relationships in the expression itself. In other words, such representations are property-structure explicit (Dienes & Perner, in press), in that their elements covary with the things they represent. Even though it is possible to imagine symbolic systems operating on representations that are compositional in a non-concatenative way (such as the representations that would be produced through Gödel numbering, for instance), such representational systems are extremely impractical, make little sense within the classical framework and violate some its basic assumptions.

These properties of symbolic representations — flexibility of access and compositionality — are of course what makes them attractive in the first place, but by the same token, it is also what makes them unsuitable for representing implicit knowledge, because there is no sense in which one can understand how such knowledge could influence processing yet remain unavailable for outside inspection. However, that is exactly what we observe in implicit learning research: Participants' performance appears to be influenced by knowledge that they do not seem to have access to. How do we reconcile the fact that a symbolic system can not represent implicit knowledge with the empirical facts? How do we turn a representation such as a production rule into an implicit representation?

Some possible answers to this question, such as knowledge compilation, are easily dismissed and will not be discussed extensively here2. Others, however, have been developed into full-blown theories of implicit learning, either by proponents or by critics of implicit learning. Interestingly, both kinds of answers leave intact the basic features of the classical framework. I discuss these theories in detail in the next section, but a brief outline will be helpful at this point.

A first strategy to make room for implicit learning within a framework that leaves no room for the implicit is to assume that some other, completely separate part of the cognitive system is responsible for it. This is a theoretical position that I call the "shadow" theory of implicit learning, because it basically postulates the existence of a cognitive unconscious that is just the same as the familiar conscious cognitive system (i.e., it uses rule-based, symbolic, abstract knowledge), only minus consciousness (see also Searle, 1992). It is best exemplified by the work of authors such as Reber (e.g., Reber, 1993) or Lewicki (e.g., Lewicki, 1986). This position is probably still dominant today, but it has long been the object of severe attacks (e.g., Dulany, Carlson & Dewey, 1984; Perruchet & Amorim, 1992; Shanks & St.John, 1994).

These attacks collectively form a second attempt to reconcile implicit learning with the symbolic framework, and are essentially eliminative in nature: Given (1) that implicit learning can not be accommodated within the symbolic framework and, (2) that clear-cut evidence for a full-blown cognitive unconscious is rather scant, perhaps the simplest way to deal with implicit learning is to consider that it does not exist as a cognitive phenomenon. Some authors (e.g., Searle, 1992)

have therefore claimed that implicit knowledge, if it exists, can not be characterized as mental. Others (e.g., Shanks & St.John, 1994) have essentially proposed that learning is always explicit, and that the difference between the kind of learning exhibited in implicit learning situations and the kind of learning exhibited in problem-solving, for instance, is in fact better captured by the dichotomy between rule-based and instance-based knowledge than by the conscious/unconscious distinction.

There is, of course, a third kind of answer to the question of how knowledge can influence performance yet not possess the properties that would enable it to be an object of representation. This answer consists of assuming that implicit knowledge is best understood as implied knowledge, in the same way as linguistic presuppositions are implied by stated expressions (Dienes & Perner, in press). From this perspective, implicit learning thus involves some form of priming through which distributional information can directly influence processing without being itself available as an object of representation. This possibility is a genuinely interesting one, not because it is successful in saving the classical framework, which it is not as I will show later, but because it is consistent with the empirical evidence and emerges naturally out of the assumptions of the connectionist framework. I will return to this argument extensively in the discussion because I believe it is the right way to characterize implicit learning.

4. Current theories of implicit learning

In this section I briefly sketch the two main contemporary theories of implicit learning. My account of each theory will undoubtedly appear somewhat simplistic, but the goal of this section is not to exhaustively characterize each framework (again, see Perruchet & Gallego, this volume, for a more empirically oriented look at each position). Instead, I merely want to set the scene for a discussion of how each approach's assumptions and methodology are in fact rooted in the classical framework.

4.1 The implicit/abstractionist framework: The "shadow" theory of implicit learning

If no one really takes the shadow theory of implicit learning literally, I believe that it nevertheless has had an underground but pervasive influence on people's thinking about the issues. In a nutshell, the shadow theory of implicit learning assumes that there is an unconscious mind that is just the same as the more familiar conscious one, only minus consciousness (see also Searle, 1992). In particular, it has been assumed that participants in implicit learning experiments are capable of acquiring abstract, rule-like knowledge implicitly (e.g., Reber & Lewis, 1977). Thus, according to the theory, the cognitive system and its shadow operate in parallel and have basically the same properties: both involve the acquisition and the processing of abstract, symbolic knowledge, albeit only the conscious system produces output available to consciousness. Over the years, different authors have ascribed additional contrasting properties to either system. For instance, the implicit system has been assumed to be faster than the conscious learning system, to operate in parallel on many variables (e.g., Berry and Broadbent, 1984), or to be resilient to the lack of attentional resources (Curran & Keele, 1993).

In a way, then, this kind of theory is an elaboration of the modularity perspective (Fodor, 1983). Indeed, Fodor's theory assumes that besides a central processor that fetches information from different databases to process the problem at hand, there is a series of encapsulated modules that can bypass central processing and produce output automatically. As Karmiloff-

Smith (1992) puts it: "Each module is like a special-purpose computer with a proprietary database" (p. 3). In its most extreme form, the shadow theory described here is merely one that assumes that there is a "general-purpose module" for implicit learning, that is, a wholly independent and encapsulated learning subsystem characterized essentially by the fact that whatever it learns is not available to consciousness. Writings by Reber (e.g., Reber & Lewis, 1977; see also Reber, 1989), by Lewicki and his collaborators (e.g., Lewicki, Czyzewska and Hoffman, 1987) or by Curran and Keele (1993) contain clear characterizations of implicit learning as involving such a separable learning system. For instance, Reber (1990) states that "Specifically, [implicit systems] (a) ought to be fairly cleanly dissociable from explicit systems, (b) should have perceptual and cognitive functions that operate largely independent of consciousness and (c) ought to show greater resiliency to and resistance to insult and injury." (p. 342). Another example can be found in Broadbent's (e.g., Berry and Broadbent, 1988) notion that learning may either be selective (S-mode learning) or unselective (U-mode learning), although this theory puts more emphasis on the processes involved than on architectural distinctions.

How did such an unsatisfactory account of implicit learning come to emerge? As indicated in the previous section, I believe that it came about as a reaction to mainstream theorizing during the 1970s. Indeed, the classical framework of cognition was the dominant metaphor for information processing when Reber introduced the notion that learning could proceed without awareness (e.g., Reber, 1967), and still exerts considerable influence today. As I described in section 3, this framework is based on the use of symbol manipulation carried out in rule-following systems — the kind of goal-directed, fully conscious activity participants engage in when solving complex formal problems (Newell & Simon, 1972). Hence Reber's fascination with participants' apparent ability to acquire and use information about material generated from finite-state grammars without being able to tell him what the rules were. Implicit learning research was thus starting to produce an altogether different picture of cognition, one that made the radical assumptions that not all that is processed is available for conscious inspection, and that some kinds of learning can proceed incidentally and not as result of hypothesis-testing, for instance. As more and more evidence for the existence of implicit learning accumulated, the problem became one of knowing how to reconcile the empirical evidence with a metaphor of cognition that leaves no room for the implicit. This is where theories such as the implicit/abstractionist theory emerged: Reber interpreted his data as evidence that a completely separate learning, representational and processing system was at play: The cognitive unconscious. Crucially, this view leaves intact the idea that cognition is essentially about symbol manipulation.

This position has been attacked recently, though not directly, by Searle (1992). In "Rediscovery of the Mind" Searle argues that if I have a rule to perform some task, that is, a symbolic representation of what action to take when faced with some stimulus in some task context, then there is nothing that in principle would prevent me from reporting this rule. This position automatically rules out unconscious rules because rules are, by definition (i.e., because they are symbolic), accessible to consciousness. Thus if a system behaves in a rule-like fashion but is unable to report the rules that it uses, then it probably does not have rules at all, it merely has the appropriate wiring to perform the task, just in the same way as a thermostat behaves in a rule-like manner without having rules. The relevant knowledge is merely hard-wired into the thermostat. Likewise, we have no access to how our visual system computes color information because this process is hard wired in the neural structures that make up our visual system.

This is an interesting argument, but I think Searle is throwing the baby away with the bath water. If I agree that we do not have unconscious rules, I also believe that there is a way for knowledge to be both mental and implicit, not in principle, but with respect to some particular context. Searle himself seems to recognize this possibility when he admits that he has "inadvertently arrived at a defense [...] of connectionism" (p. 246). I will return to this point extensively in section 6.
Another line of attack on the notion that implicit learning takes places in a "cognitive unconscious" consists of eliminating the implicit from implicit learning. This approach, which is currently very influential, is the object of the next section.

4.2 The explicit/instance-based framework: Implicit learning does not really exist

If Searle's rejection of the notion of a cognitive unconscious is based on philosophical arguments, it is empirical arguments that form the core of the explicit/associationist theory of implicit learning. Two main ideas are important in this framework. First, this position, perhaps best embodied in the work of authors such as Shanks and St.John (1994) or Perruchet and Gallego (this volume), rejects the idea that implicit learning is based on rule abstraction. Historically, as Perruchet and Gallego (this volume) point out, it is Brooks (1978) who first introduced the idea that participants' sensitivity to the deep structure of the training material in artificial grammar experiments does not necessarily imply that they have induced rules. Instead, Brooks argued, one can understand their performance just as well by assuming that they have acquired instance-based representations. Hence, according to this view, implicit learning may still well be implicit, but there are no grounds to support the notion that anything like rules are acquired by participants.

If participants acquire instances instead of rules, however, most of the existing techniques for explicit knowledge assessment need to be reconsidered. Indeed, techniques such as verbal interviews or other tests that specifically probe for knowledge about rules, now fall short of their intended purposes because they measure knowledge that participants neither possess nor need in order to perform successfully. Likewise, most of the existing dissociations between performance and explicit knowledge also need to be reevaluated. This is the second idea that characterizes the explicit/instance-based framework: Far from indicating the existence of a separate implicit learning system, observed dissociations between implicit and explicit performance are merely indicative of the fact that our measures of implicit and explicit knowledge are deficient. If one is careful enough to devise tests of explicit knowledge that actually probe participants for the knowledge they possess (i.e., instances rather than rules), then many of the observed dissociations between implicit and explicit knowledge simply vanish. Implicit learning, in this perspective, is more of an artefact than a real phenomenon.

These ideas have been brilliantly embodied by Shanks and St.John's information and sensitivity criteria (Shanks and St.John, 1994). The information criterion requires that tests of awareness should demonstrably tap on the same knowledge than the knowledge needed to support performance in the corresponding implicit test. The sensitivity criterion requires that tests of awareness should demonstrably be sensitive to all of a subject's conscious knowledge. Even though there are many reasons to believe that it would be impossible for any test of awareness to be simultaneously exhaustive and exclusive in the way required by these criteria (see Jiménez, Méndez & Cleeremans, 1996, for a discussion), Shanks and St.John's analysis made it clear that many dissociations that had been considered as solid demonstrations of implicit learning were in fact unwarranted.

To summarize, the substance of this argument is inspired by the competence/performance distinction: People do in fact possess explicit, manipulable, symbolic, compositional representations of the relevant knowledge, but our tests of this knowledge are inadequate. Hence the dissociations that we observe between performance and explicit knowledge are misleading, and there is no need to appeal to the cognitive unconscious in order to understand the data.

What is the alternative, then? Shanks and St.John (1994) have proposed (along with others, see for instance Perruchet & Amorim, 1992) the following account of implicit learning: Participants in typical implicit learning situations acquire a single database of potentially fully conscious, fragmentary, instance-based information about the stimulus material. This knowledge is demonstrably sufficient to account for participants' performance. Independent mechanisms have access to that knowledge to sustain task performance on the one hand, and explicit reports on the other hand. Usually, associations are observed between performance and explicit reports. In cases where dissociations are observed, they stem from failures of the tests of awareness to be either sensitive enough (hence failing the sensitivity criterion) or to ask participants about what they know, that is, instances rather than rules (hence failing the information criterion).

It is clear that this account of implicit learning basically does away with it. In this framework, implicit learning is simply not implicit in any interesting way, but merely because our tests of explicit knowledge are poorly designed. This reasoning in fact prompted Shanks and St.John (1994) to propose to abandon the conscious/unconscious distinction in favor of the rule-based/instance-based distinction, and led them to conclude that "human learning is almost invariably accompanied by conscious awareness" (p. 394). Shanks and St.John therefore adopt a position that Reber (1990) dubbed the "consciousness stance".

To be fair, my account of this framework is somewhat caricatural, in that recent work by authors such as Perruchet and Gallego (this volume) for instance, make it clear that one does not necessarily need to assume that the explicit instances that participants produce in direct tests of such knowledge are causal in determining performance on corresponding indirect tests. But if performance in implicit learning tasks is neither based on implicit rules, nor on explicit instances or fragments thereof, what is it based on? I return to this essential question in section 6.

5. Principles for Implicit Learning

In the previous section I have attempted to show how both proponents and critics of implicit learning have proposed theories that are consistent with the classical (i.e., symbolic/modular) framework of cognition. The implicit/abstractionist framework assumes the existence of an autonomous, but still classical (i.e., symbolic and rule-based) system. The explicit/instance-based framework asks us to imagine that learning is driven essentially by the explicit and incremental memorization of instances, and that these instances are causal both in determining performance and reports of explicit knowledge.

The goal of this section is to show how the empirical methods and concepts that characterize both theories also find their roots in central assumptions of the classical framework. I show how these methods often turn out to be unsatisfactory in that they sometimes fail to adequately

enable us to capture the complexity of the phenomena they are meant to help explore.

Psychology in general is confronted with two problems, as illustrated in Figure 1. The first problem is to bridge the gap between phenomenology and behavior, that is, to define and identify the measurable behaviors that correspond to a given phenomenological concept. The second problem is to identify the cognitive processes and representations that produce the observed behavior.

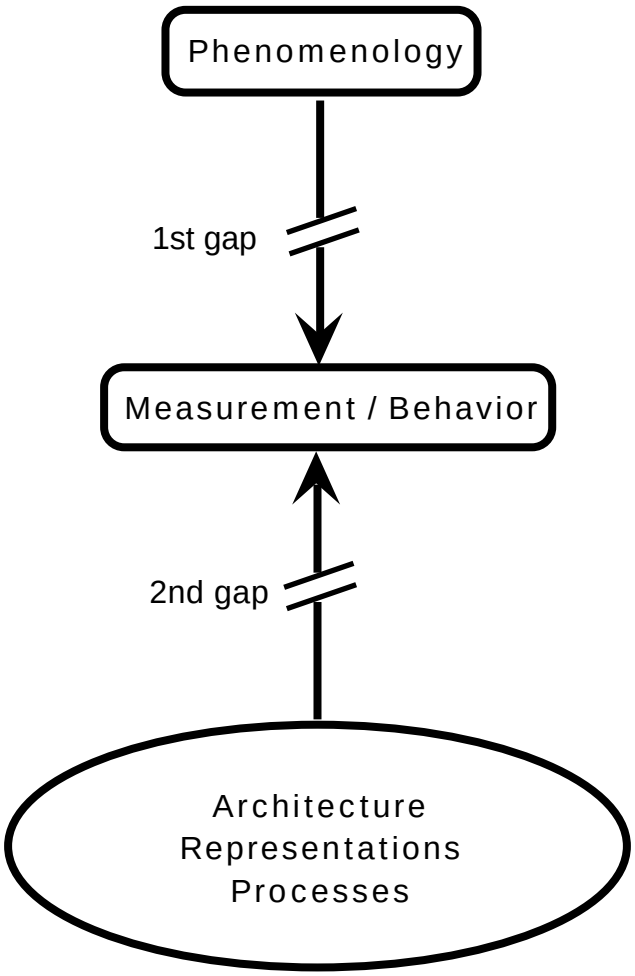

Figure 1: An illustration of the main methodological issues that face implicit learning research: How to bridge the gap that separates phenomenology from behavior and its measurement, and how to bridge the gap that separates behavior from the system that produces it.

The difficulty with bridging the first gap is that concepts such as consciousness or implicitness are vague enough that they may have multiple determinants and that it is not clear how to best operationalize them. Much of the history of research on implicit learning can be seen as attempts to move from phenomenological definitions towards objective definitions of implicit learning. However, even if we agree on how to best operationalize a given phenomenological concept so that it has a clear behavioral correlate, we are still confronted with the problem of making inferences about the causes that underpin the observed behavior.

A first problem here is that a single observable behavior may have several underlying determinants. Our environment offers many examples that illustrate this simple point. For instance, my television may fail to operate for a number of different reasons: It may be unplugged, the batteries in the remote control may be out, a fuse may be blown, the cables connecting the board to the tube may have become loose, and so on. At a given level of description, radically different causes result in the same symptom. Any complex system with many interacting components organized at different levels of description and that produces some behavior that can be defined at a gross level is bound to exhibit this "many to one" relationship between causes and effects. Thus, the symptom that the TV set is not working, if well-defined, is nevertheless coarse enough that it is not surprising to find that many different causes can be responsible for it. Likewise, knowledge may be implicit for different reasons that have little to do with each other. Whittlesea and Dorken (1993) identified four different reasons: First, one is only aware of some knowledge if the task draws attention to that knowledge. Second, the incidental learning conditions that are typical in implicit learning experiments tend to mask the fact that knowledge acquired incidentally will be relevant later on. Third, one needs some kind of theory about how knowledge that one has is actually relevant to the task at hand in order to verbalize that knowledge. Fourth, knowledge learned with awareness may be expressed as an unconscious influence rather than as an explicit act of remembering. This may be complicated in cases where the relevant knowledge is inherently distributed, as is often the case when participants are exposed to many exemplars over training. There may be additional reasons as well. For instance, knowledge acquired in tasks that essentially involves motor responses may be hard to express verbally because the knowledge is represented in a code that is not easily amenable to verbal description, or participants may lack confidence and metaknowledge when learning about inherently noisy and complex domains.

Now, a skilled technician could certainly open up the television and precisely identify the cause of the failure. However, she can only do so because she has a detailed theory of how the television work. The problem with psychology is that we do not have similarly detailed theories about the mind yet. This is the second gap. Bridging it requires that we infer from behavior what the underlying architecture and mechanisms may be. Computational modeling offers some disturbing instances of how this issue can be problematic in deep ways. For instance, it appears that different computational systems may often turn out to be functionally equivalent despite being based on different processing principles. For instance, many learning systems based on exemplars turn out to be able to produce abstract behavior and to behave in a rule-like manner without encoding rules explicitly, and some authors go as far as claiming that they are not empirically differentiable (Barsalou, 1990; Goldstone & Krushke, 1994). The performance of symbolic systems based on chunking (Servan-Schreiber & Anderson, 1990) overlaps largely with the performance of the Simple Recurrent Network (Cleeremans & McClelland, 1991) in artificial grammar learning tasks (see Berry and Dienes, 1993). Dienes (in press) also compared Logan's (1988) instance-based model with a reinforcement-based connectionist model (Barto, Sutton & Anderson, 1983) in the context of process control tasks, and again found a large overlap in how well the models accounted for empirical data.

These problems are extremely difficult ones that do not go away when one switches from one metaphor of cognition to another one. This does not mean, however, that our basic assumptions about how the cognitive system works are neutral with respect to how we think about the relationships between phenomenology, behavior, and the cognitive system. On the contrary, I believe that such background, tacit assumptions have a large impact on the conduct

of research. This issue is the focus of the rest of this section. I would like to illustrate how the classical framework relies tacitly on a cluster of simplifying assumptions that together can perhaps be described as the "assumption of direct mapping". This assumption can take many guises but basically states that there is a direct and transparent relationship between observable patterns of behavior and the internal representations and processes that produce them. As I will show below, this assumption can be applied to different kinds of descriptions of the relationship between behavior and the cognitive system, for instance to descriptions of the functional architecture or to descriptions of the internal representations of the system.

Unpacking the various components of the assumption of direct mapping and keeping them separate is not an easy task, but I will attempt a simple sketch in the following paragraphs. This sketch also provide a road map to the rest of this section, in which I have attempted to formulate a set of largely overlapping principles that I believe are important for our understanding of implicit learning. Some of these principles are metatheoretical in that they concern epistemological issues; others are grounded in computational or empirical issues. Each principle can be taken as the counterpoint to some particular aspect of the assumption of direct mapping, even though some of them have a more general reach. Taken together, I believe these principles provide an alternative framework with which to think about implicit learning.

First and foremost for the purposes of this paper is the fact that in the symbolic/modular framework, knowledge in general is assumed to be represented in the form of static and symbolic representations stored in some database, perhaps as a linked list of items or as a decision tree. Representing knowledge symbolically requires a number of additional assumptions that are problematic with respect to implicit knowledge. For instance, to do something with symbolic knowledge, you need a processor that accesses this knowledge. This makes it mandatory for the structures that represent knowledge to be distinct from the structures that process this knowledge. Two assumptions follow naturally from the the notion that knowledge is represented as symbolic items in databases.

First, observing that people exhibit sensitivity to some information is naturally accounted for by assuming that the information is represented directly in their cognitive system. If you tell me your name when I ask you, it appears natural, and even necessary in this simple example, to assume that you have a representation of your name. So far, so good. Problems start to appear, however, when we generalize this assumption to more complex cases, such as those instantiated by most typical implicit learning situations. In section 5.1 I show how there may be cases where a system exhibits sensitivity to some regularity without actually having direct representations of this regularity.

Second, architectural modularity is a natural way to explain dissociations in symbol systems. The idea that knowledge consists of lists of items stored in a database (either rules or exemplars) indeed makes it straighforward to imagine that different knowledge bases may coexist within the system, and that these knowledge bases may be completely independent from each other. One can then simply attribute observed dissociations between two measurements to the involvement of distinct processing/representational modules. This in turn mandates the double dissociation method as the method of choice to explore the organization of the cognitive architecture, and both proponents and critics of implicit have relied extensively on it to attempt to either demonstrate or invalidate the existence of separable learning systems. However, as Shallice (1988) points out, this logic appears to rely on a form of explanatory inversion: "If modules exist, then ... double dissociations are a relatively reliable way of uncovering them.

Double dissociations do exist. Therefore modules exist" (p. 248). By drawing on a variety of well-known examples and arguments, section 5.2 shows how modularity may often be only functional, and how the double dissociation logic appears to be in fact unwarranted.

The double dissociation logic also places heavy constraints on experimental design. One such constraint is that to be maximally useful, it requires our measures of behavior to be sensitive and specific enough that they can safely be taken as involving only one module or process. However, the assumption that tasks can be process-pure in this way is also problematic, as numerous authors have pointed out. Section 5.3 addresses this issue in detail.

5.1 Ontological indeterminacy

This principle is really a general one about the relationship between observable behavior and the underlying cognitive determinants. The point I want to make is simply that this relationship is a lot more complex than what some of the basic assumptions of the classical framework suggest. A lot of implicit learning research seems to have been inspired by these simplifying assumptions. For instance, Dulany, Carlson and Dewey (1984) asked participants exposed to an artificial grammar learning task to underline which letters made a string grammatical or not during classification, and found that the validity of these (explicit) judgments correlated extremely well with their (assumed implicit) classification performance. Likewise, Perruchet and Amorim, (1992) observed that performance at a serial reaction time task was well accounted for by their explicit recognition memory of small fragments of the material. What do these results tell us about participants's knowledge? A simple inference, which, on the face of it, seems very natural, is to assume that participants have simply used the representations that they can report on, and hence that performance was in fact explicit. For instance, if people tell us that they recognize a chunk of the sequence that they have been trained on, then it seems natural to assume (1) that this chunk is represented as is in their memory, and (2) that their performance at the task was driven in part by this chunk. I think, however, that there are cases where such inferences are unwarranted.

To illustrate, consider the fact that connectionist networks can be often be described as obeying rules without possessing anything like rule-like representations. A very well-known example is Rumelhart & McClelland's (1986) model of the acquisition of the past tense morphology. In the model, not only are regular verbs processed in just the same way as exceptions, but neither are learnt through anything like processes of rule acquisition. This observation is nothing new, but it has profound consequences because we often tend to ascribe rules to systems that obviously do not have any rules, like a thermostat. The point is that the thermostat does not have rules, it only has the appropriate wiring, just like connectionist networks only have their pattern of connection weights. Participants who exhibit rule-like behavior likewise cannot be assumed to possess any representation of the rules at all. Hence, observing sensitivity to some regularity does in no way imply that the regularity itself is represented within the system as an object of representation. It does not even imply that the regularity is represented as such in the system but somehow not accessible at this particular time, it only means that the system is sensitive to the regularity. In a way, the point I am making here is similar to Shanks and St.John's information criterion (Shanks & St.John, 1994): It is just as wrong to assume that because an artificial system exhibits rule-like behavior, it possesses rules as it is to assume that because people exhibit sensitivity to a rule system, they must have a representation of the rule system that is similar to the actual rules used to generate the stimulus material. Instead, Shanks & St.John (1994) proposed, one can account just as well for performance under the

assumption that participants merely acquire much more elementary information, such as memory for small fragments of the stimulus material. And indeed, participants can be shown to recognize or to be able to produce such small fragments of the stimulus material. This finding has then been interpreted as evidence that their behavior is driven by the processing of these small fragments of knowledge instead of being driven by implicit rules.

However, just as for rules, my point about representational indeterminacy also applies to memory of small fragments of knowledge. It is not because I can show that you possess a fragment of knowledge about something that this fragment of knowledge is causal in producing your behavior. It only makes sense if one also assumes that performance during the task is actually based on the processing of these small fragments of knowledge. Some computational models, however, appear to be neither based on processes of rule abstraction nor on memorization of instances.

For instance, consider Elman's (1990) original work on the Simple Recurrent Network (Elman, 1990). In several simulations, Elman demonstrated that the Simple Recurrent Network (henceforth, SRN) can learn to predict each successive item in sequences of items that it is presented with one item at a time. In one simulation, Elman generated a long sequence of letters by (1) randomly combining three consonants (b, d, and g) and (2) by replacing each consonant by a group of letters according to the following rules: "b" was replaced by "ba", "d" was replaced by "dii", and "g" was replaced by "guuu". From the point of view of a system that is attempting to predict each element of the expanded sequence, the consonant elements are completely unpredictable since their order was randomly determined, but the vowel elements are fully predictable, because both their identity as well as their number are determined based on which consonant just occurred. After training on this material, the network learnt this regularity perfectly, and produced low prediction error for all the vowel elements, but high prediction error for the consonant elements. For instance, the network cannot predict well the "g" in "guuu" but once "g" has been presented, the network shows good performance in predicting each of the three "u"s. Thus, a plot of the error as it varies over a few successive elements shows the typical sawtooth pattern that would in other contexts be interpreted as an indication that the system has chunked the input in small fragments corresponding to the expanded subsequences of the stimulus material. However, the chunks are only in the eye of the beholder, that is, they are purely functional. Indeed, the network's internal representations are very much unlike chunks. On the contrary, they are graded, distributed, and, if compositional, only in a non-concatenative way. Analyzing such internal representations may sometimes reveal well-defined clusters that could be interpreted as chunks, but this interpretation would only be descriptively correct, because the network never actually retrieves or processes representations that can be described as chunks.

Another example of how the relationship between observable behavior and the underlying processes can be inconsistent with the assumption of direct mapping is McClelland's (McClelland & Jenkins, 1990) account of the development of performance at the balance beam task. Lack of space prevents a full treatment of this example, but in a nutshell, McClelland showed how continuous internal changes can result in abrupt stage-like transitions at the level of observable behavior. Therefore, observing stage-like performance shifts in no way entails similar abrupt changes in the underlying internal representations or learning mechanisms.

All these examples show how assumptions of representational transparency are unwarranted and misleading: Unfortunately, evidence that a given system is sensitive to some regularity turns out to tell us little about how this regularity is actually represented by the system. To be

fair, the situation is probably not as bleak as depicted in the previous sentence, in that inferences about the relationship between behavior and the causal mechanisms that are responsible for it are strongly dependent on the level of description at which these inferences are made. Different theories may be equivalent at some level of description, but separable at more detailed levels of description. Finally, it is also important to stress that the examples given in this section are merely demonstrations that in some cases, the relationship between observable behavior and internal representations can be much more complex than expected based on the simple assumption of representation transparency. These demonstrations, however, do not necessarily preclude simpler relationships.

5.2. Functional Modularity

One of the basic inference tools in many fields of psychology is that double dissociations between two performance measures are caused by dissociated underlying determinants, for instance, distinct processing modules. The basic experimental logic consists of comparing performance on two tasks A & B under different conditions. A single dissociation is obtained if for instance performance on task A is better in condition 1 than in condition 2, while performance on another task B remains unchanged. In other words, the experimental variable that defines the differences between conditions 1 and 2 selectively affects one performance measure but not the other. Many examples of such dissociations can be found in the implicit learning literature. For instance, Berry and Broadbent (1984) reported that participants's ability to control a simulated system benefited from practice at the task, whereas their ability to answer a questionnaire about the system remained unchanged. A double dissociation obtains in cases where (1) a single dissociation obtains and (2) another set of conditions can be identified that reverse the relationship between performance on the two tasks. In other words, two experimental variables have opposing effects on the relationship between the two performance measures. For instance, Berry and Broadbent (1984) also showed that detailed instructions about how to control the system resulted in improved questionnaire performance yet left task performance unchanged. This kind of double dissociation is "uncrossed" in that if each experimental variable appears to selectively influence performance on one task, it leaves performance on the other task unchanged. A stronger type of double dissociation — a crossed double dissociation — would consist of observing that each variable has opposite effects on each task. With the exception of a study by Hayes and Broadbent (1988) that has failed to be replicated so far, such a pattern of results has never been observed in implicit learning situations.

The dissociation logic spelled out in this previous paragraphs has thus often been used by proponents of implicit learning as a tool to establish its existence. The reasoning that underpins such a strategy is simply that if a given manipulation appears to specifically affect one dimension of performance but not another, then there must be some specific underlying component of the cognitive system that is responsible for the first dimension of performance but is not involved in the other. Likewise, in neuropsychology, the observation that some patients suffer from a specific deficit on some task A while exhibiting intact performance on another task B has often been used as an argument to defend the notion that task A involves different processes than task B, particularly when the reverse pattern can also be shown to exist in other patients. For instance, Knowlton, Ramus and Squire (1992) have defended the position that artificial grammar learning involves processes other than those involved in explicit recollection because amnesic patients can perform well on a grammaticality judgment task despite being severely impaired on recognition as compared to normal participants.

By the same token, the fact that no crossed double dissociation has ever been satisfactorily obtained in implicit learning research has often been used by other authors (e.g., Shanks & St.John, 1994) as an argument to deny the existence of implicit learning as an independent and autonomous process. One typical strategy has been to claim that the measurements themselves are not appropriate. The dissociation logic indeed requires measures to be equally reliable and equally sensitive to the knowledge they are meant to measure for the experimental method to be valid. No conclusions about the relationship between performance and awareness can be reached, for instance, if one can demonstrate that the measure of awareness one uses fails to be sensitive enough to the relevant knowledge.

However, both positions seem to share an assumption that does not necessarily hold, that is, that the implication of observed dissociations is that the underlying cognitive subsystems are themselves dissociable. There are many reasons to question this assumption on logical or methodological grounds, but one of its most pernicious effects may be that it tends to perpetuate the notion that it makes sense to think about the issues in terms of separable or non-separable modules. Does it make sense to infer from the observation of double dissociation that independent modules are at play? Dunn and Kirsner (1988) developed a compelling argument that double dissociations are in fact insufficient to establish that two processes are independent. They showed that systems that consist of only one processing module can nevertheless produce crossed double dissociations under some circumstances, and suggested to use an alternative method which they call "reverse association". A reverse association obtains when a "pattern of association between two tasks, either positive (monotonically increasing) or negative (monotonically decreasing) is reversed in one pair of conditions relative to another pair" (p. 98). Reverse dissociations, unlike crossed double dissociations, are completely incompatible with a single underlying system. Using the reversed association method is more complex than working within the double dissociation logic, (e.g., it requires three experimental conditions rather than the usual two) but it is nevertheless surprising that so few studies have attempted to put it to work.

Other authors have appealed to theoretical and simulation work to call the dissociation logic into question. Shallice (1988), for instance, describes a number of non-modular architectures that would nevertheless produce double dissociations. To take just one of his numerous examples, damage to a specific portion of the retina would selectively impair processing of visual input at this location, yet it is clear that the retina is a continuous processing space.

Plaut (1995) explored these issues in the context of cognitive neuropsychology. Standard neuropsychological interpretations of clinical data rely heavily on an assumption that Farah (1994) describes as the "locality assumption", and which basically states that the cognitive system consists of a collection of functionally specialized processing modules that are structurally independent from each other. Double dissociations then receive seemingly natural interpretations: Damage to one specific module of the system results in deteriorated performance on tasks involving the function supported by the damaged module but has no effect on performance involving functions supported by other modules. Plaut (1995; see also Farah, 1994), however, proposed a radically different interpretation of double dissociations by showing how a connectionist network can exhibit functional double dissociation despite not being organized in architecturally distinct processing modules at all.

Plaut's argument rests on simulation studies conduced by Plaut and Shallice (1993). Plaut and Shallice's goal was to account for observed double dissociations between concrete and abstract

word reading exhibited by so called deep dyslexic patients. For instance, patient PW, described by Patterson and Marcel (1977) could only pronounce 13% of the abstract words (e.g., "truth") he was presented with. Concrete words (e.g., "table"), in contrast, elicited 67% of correct pronunciations. Hence the patient exhibits a single dissociation between concrete and abstract word reading. In and of itself this finding is not sufficient to conclude that concrete and abstract word reading are subserved by distinct processing modules, as abstract words could merely be more difficult to pronounce than concrete words, for instance. However, other patients can be shown to exhibit the opposite dissociation, that is, they perform better on abstract words than on concrete words. For instance, patient CAV (Warrington, 1981) correctly pronounced 36% of concrete words, but 55% of abstract words. As Plaut (1995) indicates, patients PW and CAV together exhibit a double dissociation. It is therefore tempting to conclude that abstract and concrete words are processed by separable underlying processing modules. However, Plaut and Shallice (1993) showed that this conclusion is unwarranted, in that it can be accounted for by systems that do not consist of such separable processing modules. To demonstrate this point, Plaut and Shallice (1993) explored the performance of a connectionist model of reading when lesioned in different ways. The network was assigned the task of producing the phonological representation of words when presented with their orthographic representation. Plaut and Shallice's model has a complex architecture, but basically consists of two interacting components linked together in a single processing pathway. First, orthographic inputs are mapped onto semantic units, which are meant to enable the network to capture the functional semantic similarities and differences between the various words of the corpus. Each unit in this pool of units corresponds to a semantic feature.

Second, the semantic units are all connected to a set of output units representing the phonological features of the words. An important aspect of this network is that unlike standard back-propagation networks, processing is fully interactive and involves recurrent connections both within and between (some) pools of units. As a result, activations can change within a single trial. To ensure that the network settles into stable patterns of activity, the pools of units corresponding to the semantics and to the phonology of the words were each connected to a separate pool of so-called clean-up units. During processing, these clean up units interact with and influence the activations of the semantic and phonological units, and in so doing force them to converge towards stable patterns of activity (i.e., attractors). After training with the back-propagation through time algorithm (Rumelhart, Hinton & Williams, 1986), the network can learn to pronounce each of the 40 words of the corpus.

There is thus nothing in this network that differentiates between abstract and concrete words in terms of specific processing components. The only feature that differentiates abstract from concrete words is the fact that, based on independent analysis of the semantic features associated with them, concrete words were represented by activating an average of 18.2% of the 98 available semantic features, whereas abstract words involved only 4.7% of these features.

Plaut and Shallice then proceeded to systematically damage the network by randomly selecting and removing some connections from each set of connections in the network. In this way, Plaut and Shallice were able to have the network reproduce the double dissociation pattern observed with human patients: Damaging the direct connections from orthographic to semantic units resulted in better performance on concrete words than on abstract words, whereas severe damage to the connections between the semantic units and their associated clean-up units resulted in better performance with the abstract words than with the concrete words. The

explanation for this difference is complex, but the gist of it that concrete words are associated with stronger semantic attractors because they involve more intercorrelated semantic features than abstract words. For instance, things that have legs and that live on the ground will often also be capable of running. The role of the clean-up units in the network is precisely to support the micro-inferences that result from the simultaneous activation of related semantic features. Hence damage to the connections to and from these units will be more adverse to the processing of concrete words than to the processing of abstract words. Processing the latter, in contrast, depends more on direct activation of specific and more independent semantic features. The crucial point, as stressed by Plaut (1995), is that both pathways are equally involved in processing either concrete or abstract words. The double dissociation is therefore not attributable to architectural specialization, but is instead a consequence of functional specialization in the representational system of the network.

It should be obvious that the findings described above have crucially important implications not only for neuropsychology, but also for psychology at large, because the double dissociation logic has been widely used in all fields of cognitive psychology. This method now appears to be flawed, based on both logical and computational arguments. Just as there are many clear cases where a particular function is dependent on the operation of a specific component of the cognitive system, there are also instances where a non-modular processing system can exhibit a double dissociation. Therefore, the notion that implicit learning entails the existence of an independent subsystem needs to be taken with extreme caution. By the same token, the fact that dissociations can be produced by a single system subtracts nothing from the fact that these dissociations may nevertheless be functionally real and may reflect underlying distinctions, for instance at the level of internal representations. Hence the critical position favored by some authors in the implicit learning field, namely that a single database of potentially conscious information drives behavior, may be accurate at some level of description, but completely fails to characterize functional distinctions that may exist within the single system, and be responsible for the observed dissociations.

5.3. Multiple-process tasks

Working with the double dissociation logic ideally requires that one can identify tasks that involve only one component of the underlying cognitive system. Just as neuropsychologists hunt for pure cases — patients whose impairment is limited to a single component of the system (Plaut, 1995); psychologists hunt for pure tasks. Dunn and Kirsner (1988) described this strategy as involving an “assumption of selective influence”, that is, the assumption that each experimental variable selectively affects a single process, and that each process contributes to a single task. However, as Dunn and Kirsner (1988) point out, this assumption is a very strong one and is rather unlikely to be correct, as even elementary tasks appear to involve many processes with largely unknown properties. It seems therefore highly unlikely that one can identify tasks that exclusively involve conscious or unconscious knowledge. Both proponents and critics of implicit learning, however, have often embraced the view (most often tacitly, as a simplifying assumption) that one can identify tasks that exclusively involve conscious or unconscious knowledge. For instance, Shanks and St.John (1994) claimed that valid demonstrations of unconscious learning should be based on dissociations between measures of implicit learning and awareness that satisfy both of their information and sensitivity criteria. However, there are many reasons to doubt that any task could satisfy both criteria (see Jiménez et al., in press, for a discussion). Indeed, authors such as Reingold and Merikle (1988; see also Merikle & Reingold, 1991) have argued that it may be impossible to identify a single measure that is simultaneously (1) exhaustively sensitive to the relevant contents of awareness and (2)

exclusively sensitive to this knowledge, because we have no way of ascertaining that tasks are process-pure, and because we do not yet have a clear theoretical understanding of awareness. Hence, instead of requiring that absolute criteria of awareness be used, Reingold & Merikle (1988) suggest that a more productive strategy may be one that consists of comparing the sensitivity of various measures of the same relevant conscious information. They start by assuming that discrimination tasks in general may involve both relevant conscious information as well as some kind of unconscious sensitivity. Thus, no measure is likely to involve either kind of knowledge and processing in isolation. However, a given measure may be characterized as a direct or as an indirect test of the relevant knowledge depending on the relationship between the discrimination that it requires and the definition of the task that participants are instructed to perform. For instance, recognition is a direct test of subject's ability to discriminate between old and new items when they are instructed to perform precisely this task. The old/new distinction, however, can also influence performance in other tasks: Merikle and Reingold (1991) have shown that judgments about the visual contrast of stimuli are affected by whether these stimuli had been presented before or not. In this case, the visual contrast judgment task would be an indirect test of the old/new distinction. Comparing similar direct and indirect measures of the same discrimination could thus be a way to enable us to determine whether performance is influenced by unconscious determinants. However, to do so, we need to make assumptions about their relative sensitivity to conscious knowledge. Reingold and Merikle propose that we make the following single assumption: Direct tests of a given discrimination should not be less sensitive to conscious, task-relevant information than comparable indirect tests. Thus, all other factors being equal, if participants are instructed to respond to information that is available to consciousness, then their use of this knowledge should not be worse than in cases where they are not directly required to use it. A straightforward implication of this assumption is that whenever an indirect measure shows greater absolute sensitivity to some relevant knowledge than a comparable direct measure does, one can conclude that this knowledge is not conscious, given that conscious knowledge alone could not explain the advantage observed in the indirect task. From this perspective, then, the most important thing one should worry about when comparing performance on implicit and explicit tasks is not whether they are pure enough — they can never be — , but whether the tasks are comparable direct and indirect measures of the same discrimination. Analyzing the implicit learning literature from this perspective is rather disturbing, because it appears that most existing tasks do not actually comply with these requirements (see Jiménez et al., in press, for a full analysis). For instance, process control (typically thought to involve implicit learning) is in fact a rather direct and explicit test of knowledge about the system. Question answering or recognition, which in this context are typically taken as measures of explicit knowledge, are also direct tests of knowledge about the system. Similar arguments apply for the tasks typically used in artificial grammar studies, such as grammaticality judgment as compared to recognition (both tasks are again direct tests). Hence in all these cases, one is comparing performance on several different direct tests of some knowledge, whereas in fact one should be comparing performance on comparable direct and indirect tests. As a result, existing associations or dissociations between implicit and explicit tasks are just as likely to reflect differences in the task contexts than they are likely to reflect differences on the conscious/unconscious dimension.

Working with tasks that are not process-pure therefore requires using new methods to assess knowledge. Over the past few years, such methods have started to make a foray in the implicit learning field. Most of them have been imported from implicit memory research, such as Merikle and Reingold's framework (1991) or Jacoby's (1991) process dissociation procedure.

Others, such as Dunn and Kirsner (1988)'s reversed association method, have not yet started to make an impact. The gist of all these methods is to enable us to take into account the fact that performance at a given task has multiple determinants. I will not review these methods in detail here, but merely give an example of their application taken from research by Jiménez et al. (in press). Jiménez et al. (in press) explored the relationship between reaction time performance and explicit knowledge as revealed through a subsequent generation task. The reaction time task was similar to Cleeremans and McClelland's (1991) situation in involving sequential material generated based on a probabilistic finite-state grammar. Through detailed partial correlational analyses (which controlled for knowledge expressed through the generation task) of the relationship between performance at the reaction time task and the statistical structure of the stimulus material as defined by the the probability of appearance of each stimulus, Jiménez et al. (in press) showed how some knowledge appears to be exclusively expressed through the reaction time task, that is, indirectly. Because the only difference between the reaction time task and the generation task is whether or not participants are asked to explicitly use their knowledge, Jiménez et al. (in press) interpreted their dissociation result as evidence for unconscious learning. Interestingly, this dissociation result obtained even though global comparisons between reaction time and generation performance suggested an association between the two measures. These surprising results thus mandate that performance be evaluated in great detail, on a trial-by-trial basis, instead of globally, as they most often typically are.

5.4. Graded and dynamic dimensions

Science seems to love dichotomies, and for very good reasons. Dichotomies are easy to think with and easy to describe. But are they right? In many cases, the answer is probably yes. However, a strong case could just as well be built about the converse proposition, namely that some dimensions are just not dichotomous. Reber (1993), who has often been cast as taking up the radical position that there may exist an unconscious learning system that is fully independent from the more familiar conscious one, nevertheless recognizes the continuous nature of many dimensions of behavior and appropriately calls the tendency to use dichotomies as the "polarity fallacy" (p. 31).

As I have argued elsewhere (Cleeremans, 1994, 1995), computational frameworks such as connectionism make it very clear that alternatives to dichotomous characterizations exist and that they often provide better accounts of the data. A convincing example of how this may be so is provided by interpretations of the Stroop interference effect (Stroop, 1935, see also Glaser & Glaser (1982) and its implications regarding automaticity. In the Stroop paradigm, participants are asked to perform one of two tasks: either read a word aloud, or name the color of the ink that the word is printed in. The difficulty of the task is that the words can be the names of colors. One can thus construct conflict stimuli in which the words and the color of the ink that they are printed in are different (for instance, the word GREEN printed in red ink), congruent stimuli in which the words and their ink colors agree, and neutral stimuli. The measure of interest is how fast participants are able to perform the tasks with the different stimuli. The typical findings are threefold. First, reading in general is faster than color naming. Second, ink color has no effect on the time it takes to read a word. Third, in color naming, conflict stimuli produce a slowdown compared to congruent or neutral stimuli. To summarize, it appears that reading is not affected by the presence of conflicting information, whereas color naming is. These results have typically been interpreted by suggesting that color naming is a controlled process whereas word reading is an automatic process. Thus, automatic processes are described as fast, involuntary, encapsulated or modular (that is, not susceptible to interference

from other processes), and impervious to the lack of attentional resources, whereas controlled processes are relatively slow, under voluntary control, and require attentional resources.

Thus, the vast literature on automaticity at some point resembled what one observes today in the implicit learning field. Automatic and controlled processes were described in terms of lists of specific properties, and countless experiments were designed to assess whether it is possible to establish that a given process is entirely automatic. Crucially for my argument, automaticity was described as a dichotomous and binary property. A given process, thus, was thus thought to be either automatic or controlled.

Cohen, Dunbar & McClelland (1990), along with others before them (e.g., Logan, 1980; Kahneman & Chajczyk, 1983; McLeod and Dunbar, 1988) attacked this position and claimed that automaticity is really a continuous dimension. Illustrative empirical arguments can be found in a study by McLeod and Dunbar (1988), in which participants were placed in a modified version of the Stroop paradigm. McLeod and Dunbar asked their participants to learn to use color names as the names for arbitrary shapes and trained them on this shape-naming task for a very large number of trials spread over 20 days. At different points during training, participants were asked to perform a Stroop task that involved shape-naming and color-naming as the relevant dimensions instead of the usual word reading and color-naming. They found that the names of the shapes tended to interfere more and more with color-naming as training progressed. Thus, a shape's name that conflicts with the color that the shape is printed in would have no effect on color naming speed early in training, but produced an increasingly large slowdown with practice at the shape naming task. Clearly, then, one can obtain continuous interference effects, the magnitude of which depends on the amount of training that each dimension has been allowed to benefit from. The modeling work of Cohen et al. (1990) showed that the effects of practice can be simply expressed as the strength of a processing pathway in a connectionist model. The model was successful in accounting for McLeod and Dunbar's data, and suggests that automaticity, rather than being dichotomous, is best thought of as a continuous dimension that is related to the relative strength of different processing pathways.

Just as for automaticity, I believe that many dimensions of cognition that play an important role in the implicit learning field (and have been the object of intense debate), such as the abstraction, or awareness, may in fact be graded and continuous rather than discrete and dichotomous. I address each in turn briefly in the following paragraphs.

Consider how connectionist networks process information. Over the course of training, a network is exposed to exemplars and trained to produce the appropriate response when presented with each exemplar. In a trained back-propagation network, the distribution of the activation vectors of internal units represents a mapping between inputs and outputs that is sufficient to compute the transfer function required to assign each input exemplar to its correct output category. It has long been known that the ability of a network to generalize to new items, that is, to correctly assign new input patterns to their correct response categories, is related to the number of internal units of the network (e.g., Hinton, 1986). Few internal units usually result in better ability to generalize because the relevant dimensions of the input domain had to be extracted and compressed over training. Hence the only way for a network with few internal units is to represent only the most general dimensions of the input. Large numbers of hidden units, by contrast, often result in poor ability to generalize because redundant representations were allowed to develop over training. For instance, Hinton (1986) trained a

back-propagation network to process linguistic expressions consisting of an agent, a relationship, and a patient, such as for instance "Maria is the wife of Roberto". The stimulus material consisted of a series of such expressions, which together described the family trees of an italian family and of a english family. The network was required to produce the patient of each agent-relationship pair it was given as input. For instance, the network had to produce "Roberto" when presented with "Maria" and "wife". Hinton showed that after training, the network had developed internal representations that captured relevant abstract dimensions of the domain, such as nationality, sex, or age. The crucial point is that the input representation contained no information whatsoever about these abstract dimensions: Each person or relationship was simply represented by activating their corresponding input unit. Further, the model generalized to new instances of specific input-output pairings that had never been presented during training (albeit in only a limited number of test cases). Thus, in Hinton's words, "The structure that must be discovered in order to generalize correctly is not present in the pairwise correlations between input units and output units" (p. 9). The model thus exhibits sensitivity to relational similarity based on the distributional information present in the input: Based on processing exemplars, the model has developed abstract knowledge of the relevant dimensions of the domain.

Examples from work on recurrent connectionist architectures such as the SRN also support the notion that training based on exemplars can nevertheless be sufficient to produce rule-like behavior and rule-like representations. For instance, an SRN trained on only some of the strings that may possibly be generated from a finite-state grammar will generalize to the infinite set of all possible instances (see Servan-Schreiber, Cleeremans & McClelland, 1991). It will sometimes develop internal representations that are organized in clusters, with each cluster representing a node of the grammar — as abstract a representation as could be. In other cases, however, the network's internal representations tend to be organized in numerous very small clusters that each correspond to one or to a few training instances (see Servan-Schreiber et al., 1991; Cleeremans, 1993; for detailed examples). The SRN has often been described as processing fragmentary information. This is descriptively correct, but it is not how things work inside the network. The network does not develop a database of subsequences that it can consult and ponder about as a result of training. Instead, as it processes each stimulus, the constraints that exist between the successive elements of the sequence are progressively incorporated in the pattern of connection weights so as to allow the network to respond better to the task demands. Note that the fact that subsequences are not explicitly represented in the network does not make it incapable of recognizing such sequences either. The network can indeed be used as a finite-state recognizer (Servan-Schreiber et al., 1991).

Therefore, connectionist networks of this kind are clearly much more than simple associators that only encode input-output correspondences based on a set of stored training examples. Indeed, as McClelland & Rumelhart (1985) suggest, depending on factors such as the number of hidden units or the structure of the training set, such networks may develop internal representations that are best characterized as storage of exemplars (i.e., many micro-associations) or as an encoding of the shared properties of many instances (i.e., a few general associations). Thus, there appears to be a representational continuum that extends from raw storage of instances to fully abstract representations, and the opposition that is often made between abstract (implicit) knowledge and fragmentary (explicit) knowledge that is at the heart of so many debates about implicit learning performance begins to fade away when one considers the way in which connectionist models represent and use information. In short, abstraction is a graded, dynamic dimension.

Do similar arguments hold in the case of awareness? This is a much more difficult question, because answering it requires one to have a theory of consciousness first. As a result, computational frameworks are almost universally silent about the issue of awareness. There are a couple of points worth stressing, however. First, the phenomenology of awareness certainly seems to suggest that it is a dynamic rather than static dimension. What I am aware of now I may be unaware of at the next moment. Early characterizations of implicit learning, however, have often tended to describe availability to awareness as a static property of knowledge, because of their reliance on separate "implicit" and "explicit" knowledge bases to account for observed dissociations. It seems clear that this is the wrong way to think about the distinction between implicit and explicit knowledge. Availability to awareness, or the property of being explicit, are dynamical properties in the sense that you can only speak of awareness of something at a particular point in time. Thus at any given point in time, I am aware of a subset of my total knowledge and unaware of the rest. In other words, the contents of awareness at some point in time are those pieces of knowledge that fall under the current "spotlight of attention". This characterization of the contents of awareness as the items that are currently active in memory has been incorporated in influential models such as Anderson's ACT* production system architecture (Anderson, 1983).

The problem with such conceptions of awareness, and this is my second point, is that they tend to describe it as an all-or-none property of knowledge rather than as a continuous dimension. The phenomenology is somewhat more complex to describe for this property of awareness, in that there are some ways in which awareness appears to be a graded dimension, and other ways in which it appears to be an all-or-none property. The main difficulty with the classical framework is that in many cases, it takes it as a starting point that availability to awareness is an all-or-none property. For instance, Ling and Marinov's (1994) model of performance in sequential reaction time tasks completely fails to provide the means of characterizing its knowledge as implicit, short of tagging it as implicit. There seems to be no room in the model for knowledge that is somewhere in between on a continuum of awareness. This is problematic because one needs to have some way of allowing this knowledge to influence performance without being available to awareness in order to understand implicit learning. One way to solve this problem is to associate an activation level to each piece of knowledge and to assume that activation needs to exceed a given threshold for the corresponding piece of knowledge to enter awareness. Provided one also allows partially activated knowledge to influence processing, this arrangement would work, but it also appears unduly artificial and arbitrary. Connectionist models, if they are no more successful in providing answers as to why we become aware of some knowledge (but see Mathis and Mozer, 1996), at least make it natural for knowledge to be graded in nature in terms of its relative accessibility. I return to this point in the next section.

The view expressed here, then, is one in which cognition is viewed as involving essentially continuous dimensions. Whatever symbolic properties come out of this essentially continuous representational system come about because of our use of language, or because of memory constraints, and are produced as a result of functional adaptation to the demands of the the environment. I believe that this perspective allows for a far more natural characterization of both implicit and explicit learning than any perspective that assumes discrete, symbolic representations from scratch.

6. Connectionism and Implicit Learning

In the previous section, I have attempted to show how current theories of implicit learning often make simplistic assumptions about the relationship between measurable behavior and the causal mechanisms responsible for it. The gist of this demonstration has been to suggest that these assumptions are in place because of the tacit adoption of the classical framework as the metaphor of mind. In this section, I would like to focus on some properties of knowledge representation and of processing in connectionist networks, and illustrate how these properties provide better natural primitives to think about implicit learning than the classical metaphor.

In what way do the assumptions of the connectionist framework differ from those of the symbolic/modular framework with respect to our understanding of implicit learning? I already addressed this issue to some extent in section 2, but it is worth going over the main points again here. Two key features are important making connectionist networks better tools to think about implicit learning than symbolic models.

The first one is that learning in connectionist networks does not involve accumulating pieces of knowledge in a dedicated part of memory. That is, learning is not necessarily driven by incremental memorization, as in symbolic/modular systems. In such systems indeed, learning always involve (1) incorporating new knowledge in a database of facts or rules, or (2) combining existing pieces of knowledge to produce more complex pieces of knowledge. Somehow, in this framework, one always thinks of cognition as a processor that runs a program that operates on representations. Crucially, this is true both for abstraction-based models and for exemplar-based accounts. By contrast, in connectionist networks, learning is thought to be a by-product of processing, and involves changing the very structures that drive processing (that is, the connection weights between units). It is easy to see why implicit learning is a problem in the first metaphor. Indeed, if one assumes that the only learning mechanism that is available is one that adds pieces of knowledge to databases, then the only way to understand implicit learning is to assume that somehow, learning is implicit, that is, that the process that adds the information to the databases produces knowledge that is not available for outside inspection.

The second crucial difference between the two frameworks with respect to the issues at hand is that knowledge in classical systems is assumed to be represented symbolically and that symbolic representations are typically compositional in a specifically concatenative way, that is in a way that explicitly preserves the elements of the representation in the representation itself. In connectionist networks, however, knowledge is not represented as discrete symbolic entities, but rather as patterns of activation that are distributed over many processing elements. Because of their distributed nature, connectionist representations are not compositional in the classical sense of the term, that is, they are not concatenative. Elements of such distributed knowledge structures can therefore influence performance directly, without first having to be extracted and interpreted, and without being represented as separate, potentially manipulable objects of representation.

What is the kind of characterization of implicit knowledge that emerges out of the connectionist framework? I believe implicit knowledge, to use Perner and Dienes's (in press) terminology, to be best characterized in the same way as linguistic presuppositions. For instance, the sentence "John payed the bill" simultaneously opens up and constrains the representations that a listener may develop of what happened. That "John payed the bill" is represented explicitly when I hear the sentence is obvious. However, it is far from clear whether the implications of "John payed the bill" are represented in the same way. For instance, the fact that "John payed the bill" is

consistent with John being in a restaurant and just having had a meal, or with John being in a shop and just having bought an item, or with John having committed some crime and now being sentenced to a prison sentence, and so on. The sentence is also consistent with John being male, human, of sufficient age to be capable of paying a bill, and so on. When we process language, we are not directly aware of all the implications of what we hear. Granted, this kind of knowledge may become explicit in the course of processing, or may be brought to awareness when we are specifically probed about it, or when we meet some inconsistent subsequent statement. But the point is that we are not aware of all the implications of the statement when we process it. However, it is undeniable that processing the statement constraints in many ways which other statements are consistent or not. Hence knowledge of which one is not directly aware of at some point in time is nevertheless brought to bear on subsequent processing. I believe that this is exactly what is happening in many implicit learning experiments. For instance, one could describe processing during a sequential choice reaction experiment in just the way I described processing the linguistic example above. Consider for instance Cleeremans & McClelland's (1991) experiments, during which subjects were exposed to 60,000 trials of a sequential choice reaction time task. In sharp contrast to the simple repeating short deterministic sequences used in the vast majority of sequence learning experiments, the stimulus material we used was generated from a probabilistic and noisy finite-state grammar. Hence, almost all permutations between elements of subsequences of any length appear during training, albeit with different frequencies. There is an infinite number of such sequences, and still thousands of them if one only considers subsequences of up to 6 elements. I find it utterly implausible to assume that participants who were merely instructed to respond to the current stimulus would somehow consciously encode and memorize all these possible subsequences and use this knowledge to explicitly prepare for the next event. Yet, the reaction time data shows exquisitely detailed sensitivity to the ensemble of constraints resulting from an encoding of all the subsequences (see Cleeremans, 1993). There is no evidence whatsoever that subjects have conscious access to this kind of distributional information about the stimulus material. Further, the fact that subjects can consciously retrieve specific instances does not tell us anything about whether these instances are what performance is based on, nor does it tell us anything about how they are used, if at all, during learning. The fact that the crucial distributional information gets represented spontaneously as a side effect of processing in connectionist networks provides a natural way of understanding why the knowledge can be "in the system" yet not be available for inspection by some other component of this system.

Is there any way similar mechanisms could be implemented in the classical framework? Characterizing implicit knowledge in this way within a symbolic framework is not impossible, but appears to produce rather implausible or else purely descriptive interpretations of the data, in that the constraints set by some piece of knowledge necessarily have to be computed somehow to have any effect on further processing. Computing them in turn entails that they be explicitly represented at some point. So for instance I may have a restaurant script that enables me to infer from a sentence such as "John paid the bill" that John was probably in a restaurant, but this information will have to be explicitly represented at some point, as an additional piece of knowledge in working memory, for instance, for it to be capable of influencing further performance.

Granted, one may claim that such priming occurs fast enough for it to fail to reach awareness, or in such a way that the primed elements remain below some "awareness" threshold, for instance, but this again appears to be an arbitrary rather than natural feature of the resulting model. In contrast, even early models such as McClelland' Jets and Sharks example

(McClelland, 1981) illustrate how several properties related to priming, such as content addressability or spontaneous generalization, emerge naturally out of the model. As McClelland, Rumelhart and Hinton (1986) state: "These properties must be explicitly implemented as complicated computational extensions of other models of knowledge retrieval, but in PDP models they are natural by-products of the retrieval process itself." (p. 31).

More recently, van Gelder (1990) has shown how the kind of functional compositionality that emerges out of connectionist systems that use distributed representations (such as most current connectionist models) has the potential to offer a radically different perspective on cognition and on awareness. Van Gelder suggests that one approach "[...] is to devise models in which structure-sensitive processes operate on the compound representations themselves without first stopping to extract the basic constituents. These processes must capitalize directly on the inherent and systematic structural similarities among the nonconcatenative representations. In such models it is not only storage that takes place in the nonconcatenative domain, but the primary processing responsible for systematic behavior as well". (p. 381). A simple example of such processing is perhaps again provided by the SRN model. Over training, the network learns to develop compact representations of the sequence it is exposed to. These representations influence processing in that, together with the current element of the sequence, they determine what the next events may be. However, individual sequence elements are nowhere to be found in the network's representations of the temporal context, precisely because these representations are not concatenative. Plaut (1995) makes the same point in the context of his model of word reading (Plaut and Shallice, 1993): "[...] in a distributed attractor network, there is noting in the structure of the system that corresponds to a word. Rather, the lexical status of a string of letters or phonemes depends solely on functional aspects of the system: How particular patterns of orthographic, phonological and semantic activity interact to form stable patterns as a result of the system's knowledge encoded in connection weights" (p. 9). This is in stark contrast with models such as Ling and Marinov's, the rules of which explicitly preserve each sequence element as an object of representation.

This opaque character of connectionist representations is both a virtue (in that it offers a principled and radically new way of understanding how cognition may emerge from processes that do not manipulate symbols) and, for an increasingly larger number of authors, a problem. For instance, Karmiloff-Smith (1992) provides convincing arguments that development must involve a process that she dubs "representational redescription", that is, an active re-representation and transformation of internal states. The challenge, from this perspective, is therefore to understand how symbol processing emerges out of first-order systems such as connectionist networks. At this point, although these models are not easily amenable to implement such processes, the many recurrent architectures that have been proposed over the past few years (e.g., Elman, 1990) nevertheless make it plain that it is possible for a network to use its own internal representations as objects of knowledge capable of further influencing performance.

7. Discussion

In this chapter I have attempted to put together many ideas that I first came in contact with through research on connectionism. Implicit learning is a field characterized by complex data that requires complex interpretations in order to attempt to answer even more complex questions. This does not make it unique. What does make it somewhat unique, however, is that the field is at a stage where theoretical statements tend to be radical but unsupported by the data.

This is likely due to the fact that we lack the methodological sophistication other fields have attained, such as implicit memory. It may also be due to the fact that the principles I spelled out earlier tend not to be best captured by traditional frameworks for understanding cognition, and that most current thinking in the field is still driven by such traditional frameworks. As I tried to show, these principles are better embodied in connectionist networks, which I believe provide far better natural primitives than other frameworks not only to describe implicit learning, but also to think about it.

Interestingly, connectionist models have often been rejected for the same reasons that experimental data about implicit learning have often been subject to controversy: Both seem to offer a picture of cognition as essentially intractable. There may be several reasons why this is so. First, connectionist models are often applied to complex, fuzzy, and large problems. By the same token, implicit learning paradigms often also involve complex, fuzzy and large problems. Thus in both cases the environment to which the system is exposed already involves a kind of complexity that is neither algorithmic nor artificial. Second, connectionist models develop solutions to these learning problems that typically involve complex, time-varying, distributed representations for which most standard analysis tools are woefully inadequate. Often, it is not clear whether a particular model has developed abstract representations of the stimulus material or not, for instance. In other cases, dissociations between two aspects of performance can be obtained within a single system, thereby shattering one of our basic inference tools. I believe that we are confronted to the same kind of issues when we try to understand implicit learning performance, and that these difficulties contribute a lot to the current controversies.

The basic problem with connectionism as a modeling tool is thus also precisely what also makes it attractive: Emergent, complex, dynamical behavior — just as with human participants! In terms of research strategy, the complexity and the variability of connectionist models are properties that should be considered as a virtue in that models that exhibit truly emergent behavior are probably rich enough to enable us to capture the complexity of the data. The drawback is that working with connectionist models often entails adopting the same methods as one uses with human participants, that is, experimentation. More and more theories based on connectionist modeling are now based on statistical analysis of the behavior of different groups of individual networks (i.e., Gibson and Plaut, 1995), and the overall strategy often consists of a dual exploration of the modeling and empirical spaces. Hence connectionism in and of itself does not help solve any of the methodological issues I raised in section 5, but instead offers new principled ways of interpreting the data patterns uncovered by traditional methods, such as the double dissociation method.

In my perspective then, the future of implicit learning involves outgrowing the current set of experimental paradigms, and most importantly, the current knowledge assessment and analytical methods we use. Both of these objectives can be attained by exploiting tools and techniques used in other related fields, such as implicit memory, computational modeling, neuropsychology, and consciousness research. Connectionism, because it departs so radically from the standard assumptions we make about cognition and seems to provide better natural primitives to think about implicit cognition, appears to offer the most interesting avenue of theoretical development.

Hence we end almost where we started: The hard questions that Reber started asking himself in 1965 are still unanswered today. As many other people, including Broadbent (e.g., Broadbent, 1992), have repeatedly pointed out, the main problem involves defining awareness and

consciousness. Before we have a clear understanding of what it means to be conscious and of what the role of consciousness may be in cognition, it would appear that efforts to determine how to best measure it are, at the very least, bound to be problematic. In this paper, I have suggested that symbol systems, because of the way they represent and process knowledge, take it as a starting point that knowledge is explicit. Connectionist models, by contrast, make it clear how knowledge can be implicit, in the sense of how knowledge can be in the system, influence processing, and yet not be available as an object of representation itself. From this perspective then, the real problem is understanding consciousness, not implicit learning.

To conclude, I do not mean to suggest that the field is going circles and not getting anywhere. Rather, I propose to sidestep the entire issue and to consider what would happen if this field were merely called “learning”.

Author Note

The author is a Research Associate of the National Fund for Scientific Research (Belgium). I thank Mark St.John and Pierre Perruchet, who provided valuable reviews on an earlier version of this paper. I also thank Alain Content, Bob French, Luis Jiménez and Tim Van Gelder, all of whom contributed critical and stimulating discussion about the issues. Portions of this chapter were adapted from Cleeremans (1994) and from Cleeremans (1995).

Notes

1. In the following, I assume that implicit learning and processing exist, that is, that it is at least phenomenologically valid to distinguish between cases where we seem to have full access to the knowledge that we use in some particular context and cases where we do not seem to enjoy such access to the knowledge that governs our behavior.

2. Implicit knowledge may be knowledge that is stored in a compiled form. For instance, production systems like SOAR (Newell, 1990) or ACT* (Anderson, 1983) assume that performance improvement in general results from the fact that previously separate production rules can become combined over training as the system detects systematic sequences of rule firings. Several rules can then become combined into a single rule that fires whenever the conditions of the first rule in the chain are fulfilled, and that outputs the actions of the last rule in the chain. Knowledge compilation provides an elegant account of why experts often find it difficult to spell out and analyze their own decisions: It is because the intermediate steps involved in solving a problem become progressively embedded in the complex rules created during learning. If this appears to provide a potential mechanism for representing implicit knowledge, it is crucial to stress that there is nothing in the compiled representations that make them in principle inaccessible to the system. For instance, the system could keep a copy of the individual rules that were combined into a single compiled rule. Or it could disassemble the compiled rule, just in the same way as it was able to compile it. In other words, there is nothing intrinsic in the nature of the compiled rules that make them inaccessible to the system's "awareness" of itself. A second problem is that if knowledge compilation makes it clear how knowledge may become less accessible and more efficient with expertise, it remains unable to account for implicit learning, that is, the acquisition of knowledge that is never explicitly represented.

References

Anderson, J.R. The architecture of cognition. Cambridge, MA: Harvard University Press.

Barsalou, L.W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T.K. Srull & R.S. Wyer (Eds.), Advances of social cognition (vol. 3): Content and process specificity in the effects of prior experiences.

Barton, A.G., Sutton, R.S., & Anderson, C.W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13, 834–846.

Bates, E.A., & Elman, J.L. Connectionism and the study of change. Center for Research on Language, University of California, San Diego: CRL Technical Report 9202.

Berry, D.C., and Dienes, Z. (1993). Implicit Learning: Theoretical and empirical issues. Hove (UK): Lawrence Erlbaum.

Broadbent, D.E. (1992). Descartes, Turing and the future: The study of machines and people that change their nature. International Journal of Psychology, 27:1.

Brooks, L.R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B.B. Lloyd (Eds.), Cognition and categorization. New York: Wiley.

Cleeremans, A. (1993). Mechanisms of implicit learning: Connectionist models of sequence processing. Cambridge, MA: MIT Press.

Cleeremans, A. (1994). Awareness and abstraction are graded dimensions. Behavioral and Brain Sciences, 17, 402–403.

Cleeremans, A. (1995). No matter where you go, there you are. American Journal of Psychology, 108, 589-598.

Cleeremans, A. & McClelland, J.L. (1991). Learning the structure of event sequences. Journal of Experimental Psychology: General 120: 235–253.

Cohen, J.D., Dunbar, K., & McClelland, J.L. (1990). On the control of automatic processes: A parallel distributed account of the Stroop effect. Psychological Review, 97, 332–361.

Curran, T., & Keele, S.W. (1993). Attentional and non-attentional forms of sequence learning. Journal of Experimental Psychology: Learning, Memory, and Cognition 19: 189–202.

Dienes, Z. (in press). Memory array models of controlling a complex system. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Dienes, Z, & Perner, J. (in press). Implicit knowledge in people and in connectionist networks. In G. Underwood (Ed.), Implicit Cognition, Oxford, England: Oxford University Press.

Dulany, D.E.; Carlson, R.C. & Dewey, G.I. (1984). A case of syntatical learning and judgment: How conscious and how abstract?. Journal of Experimental Psychology: General, 113, 541-555.

Dunn, J.C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. Psychological Review, 95, 91–101.

Elman, J.L. (1990). Finding structure in time. Cognitive Science 14: 179–211.

Farah, M.J. (1994). Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. Behavioral and Brain Sciences, 17, 43–104.

Fodor, J. (1983). The modularity of mind. Cambridge, MA: MIT Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architercture: A critical analysis. Cognition, 28, 3-71.

Gibson, E., & Plaut, D. (1995). A connectionist formulation of learning in dynamic decision-making tasks. In Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society, 512–517.

Goldstone, R.L., & Krushke, J.K. (1994). Are rules and instances subserved by separate

systems? Behavioral and Brain Sciences, 17, 405.

Glaser, M.O., & Glaser, W.R. (1982). Time course analysis of the Stroop phenomenon. Journal of Experimental Psychology: Human Perception and Performance, 8, 875–894.

Hayes, N.A. & Broadbent, D.E. (1988). Two modes of learning for interactive tasks. Cognition, 28, 249-276.

Hinton, G.E. (1986). Learning distributed representations of concepts. In Proceedings of the Eighth Annual Conference of the Sognitive Science Society. Hillsdale, NJ: Lawrence Erlbaum.

Jacoby, L.L. (1991). A process dissociation framework: separating automatic from intentional uses of memory. Journal of Memory and Language, 30, 513- 541.

Jiménez, L, Méndez, C., & Cleeremans, A. (in press). Comparing direct and indirect measures of sequence learning. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Karmiloff-Smith, A. (1992). Beyond Modularity, Cambridge, MA: MIT Press.

Kahneman, D., & Chajczyk, D. (1983). Tests of the automaticity of reading: Dilution of stroop effects by color-irrelevant stimuli, Journal of Experimental Psychology: Human Perception and Performance, 9, 497–509.

Knowlton, B, Ramus, S., & Squire, L. (1992). Intact artificial grammar learning in amnesia: Dissociation of classification learning and explicit memory for specific instances. Psychological Science, 3, 172-179.

Lewicki, P. (1986). Nonconscious social information processing. New York: Academic Press.

Lewicki, P.; Czyzewska, M. & Hoffman, H. (1987). Unconscious acquisition of complex procedural knowlwedge. Journal of Experimental Psychology: Learning, Memory and Cognition, 13, 523-530.

Ling, C.X., & Marinov, M. (1994). A symbolic model of the nonconscious acquisition of information. Cognitive Science, 18, 595–621.

Logan, G. (1980). Attention and automaticity in Stroop and priming tasks: Theory and data. Cognitive Psychology, 12, 523–553.

Logan, G. (1988). Towards an instance-based theory of automatization. Psychological Review, 95, 592–527.

MacLeod, C.M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. Journal of Experimental Psychology, 14, 126–135.

McClelland, J.L. (1981). Retrieving general and specific information from stored knowledge af specifics. Proceedings of the third annual meeting of the Cognitive Science Society, 170–172.

McClelland, J.L., & Jenkins, E. (1990). Nature, nurture, and connectionism: Implications for connectionist models of cognitive development. In K. van Lehn (Ed.), Architectures for intelligence. Hillsdale, NJ: Lawrence Erlbaum.

McClelland, J.L., & Rumelhart, D.E. (1985). Ditributed memory and the representation of general and specific information. Journal of Experimental Psychology: General, 114, 159–188.

McClelland, J.L., Rumelhart, D.E., & Hinton, G. E. (1986). The appeal of Parallel Distributed Processing. In D.E. Rumelhart & J.L. McClelland (Eds.), ParallelDistributed Processing: Explorations in the microstructure of cognition (Vol. 1, pp. 3–44), Cambridge, MA: MIT Press.

Mathis, D., & Mozer, M. (in press). Conscious and unconscious perception: A computational theory. In Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Lawrence Erlbaum.

Merikle, P.M., & Reingold, E.M. (1991). Comparing direct (explicit) and indirect (implicit)

measures to study unconscious memory. Journal of Experimental Psychology: Learning, Memory and Cognition 17: 224–233.

Newell, A., & Simon, H.A. (1972). Human Problem Solving. Englewoods Cliffs, NJ: Prentice-Hall.

Newell, A. (1980). Physical symbol systems, Cognitive Science, 4, 135-183.

Newell, A. (1990). Unified theories of cognition. Cambridge, MA: Harvard University Press.

Patterson, K.E., & Marcel, A.J. (1977). Aphasia, dyslexia, and the phonological coding of written words. Quarterly Journal of Experimental Psychology, 29, 307–318.

Perruchet, P. & Amorim, P.A. (1992). Conscious knowledge and changes in performance in sequence learning: evidence against dissociation. Journal of Experimental Psychology: Learning, Memory and Cognition, 18, 785-800.

Perruchet, P., & Gallego, J. (in press). An associative model of implicit learning. In D. Berry (Ed.), What is implicit about implicit learning?, Oxford, Englang: Oxford University Press.

Plaut, D.C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. Journal of Clinical and Experimental Neuropschology, 17, 291–326.

Plaut, D.C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. Cognitive Neuropsychology, 10, 377–500.

Reber, A.S. (1967). Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior: 6, 855–863.

Reber, A.S. (1989). Implicit learning and tacit knowledge. Journal of Experimental Psychology: General, 118, 219-235.

Reber, A.S. (1990). On the primacy of the implicit: Comment on Perruchet and Pacteau. Journal of Experimental Psychology: General, 119, 340–342.

Reber, A.S. (1993). Implicit learning and tacit knowledge. Oxford University Press.

Reber, A.S. & Lewis, S. (1977). Implicit learning: An anlysis of the form and structure of a body of tacit knowledge. Cognition, 5, 333–361.

Reingold, E.M. & Merikle, P.M. (1988). Using direct and indirect measures to study perception without awareness. Perception and Psychophysics, 44, 563-575.

Rumelhart, D.E., Hinton, G.E, & Williams, R.J. (1986). Learning representations by back-propagating errors. Nature, 323, 533–536.

Rumelhart, D.E., and McClelland, J.L. (1986). On learning the past tense of english verbs. In J.L. McClelland & D.E. Rumelhart (Eds.), Parallel Distributed Processing: Explorations in the microstructure of cognition (Vol. 2, pp. 216–271) Cambridge, MA: MIT Press.

Searle, J.R. (1992). The rediscovery of the mind. Cambridge, MA: MIT Press.

Servan-Schreiber, D., Cleeremans, A. & McClelland, J.L. (1991). Graded State Machines: The representation of temporal contingencies in simple recurrent networks. Machine Learning, 7, 161–193.

Servan-Schreiber, E., & Anderson, J.R. (1990). Learning artificial grammars with competitive chunking. Journal of Experimental Psychology: Learning, Memory, and Cognition 16, 592–608.

Shallice, T. (1988). From neuropsychology to mental structure. Cambridge, England: Cambridge University Press.

Shanks, D.R. & St.John, M.F. (1994). Characteristics of disociable learning systems. Behavioral and Brain Sciences, 17, 367-395.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18, 643–662.

van Gelder, T. (1990). Compositionality: A connectionist variation on a classical theme. Cognitive Science, 14, 355–384.

Warrington, E.K. (1981). Concrete word dyslexia. British Journal of Psychology, 72,

175–196.

Whittlesea, B.W., & Dorken, M.D. (1993). Incidentally, things in general are particularly determined: An episodic-processing account of implicit learning. Journal of Experimental Psychology: General 122: 227–248.