

Unifying Consciousness with Explicit Knowledge

Zoltan Dienes
University of Sussex

Josef Perner
University of Salzburg

In Cleeremans, A. (Ed.) The unity of consciousness: binding, integration, and dissociation.
Oxford University Press, forthcoming.

A. Abstract

In this chapter we establish what it is for something to be implicit or explicit. The approach to implicit knowledge is taken from Dienes and Perner (1999), which relates the implicit-explicit distinction to knowledge representations. What it is for a representation to represent something implicitly or explicitly is defined and those concepts are applied to knowledge. Next we will show how maximally explicit knowledge is naturally associated with consciousness. We argue that each step in a hierarchy of explicitness is related to the unity of consciousness and that fully explicit knowledge should be associated with a sense of being part of a unified consciousness. New evidence indicating the extent of people's implicit or explicit knowledge in an implicit learning paradigm will then be presented. This evidence will indicate people can be consistently correct in dealing with a context-free grammar while lacking any knowledge that they have knowledge.

1. Introduction

In this chapter we will show how an understanding of the nature of explicit knowledge helps show why consciousness should have an apparently unified character. We start by taking a representational theory of the mind, specifying how representations can have both explicit and an implicit content. When this approach is applied to what it is to have knowledge a hierarchy of ways in which knowledge can be explicit is produced. Each step in the hierarchy has relevance for the production and appreciation of a unified consciousness. Implicit knowledge - for example, the implicit knowledge produced by learning some artificial grammars - does not fully take part in this unity by virtue of its implicitness. We will end by considering a specific new experiment illustrating how knowledge of an artificial grammar can be shown to be implicit by our framework. That, at least, is the argument we now try to develop, starting from the beginning: The nature of representation.

2. The implicit and explicit content of representations

In order to be clear about what has been implicitly or explicitly represented, we need a theory of what determines a representation's content. In the past, we (Dienes & Perner, 1996, 1999, 2001a,b; Perner & Dienes, 1999) have turned to functional theories of representations, i.e. the content of a representation is determined by the functional role the representation plays. Previously we illustrated this approach with Dretske's (1988) account of representation: If A has the function of indicating B, then A represents B. For example, a bee dance has the function of indicating the location of nectar, so the dance represents the location of nectar. Here we will consider another functional theory, namely,

that of Millikan (1984, 1993), to show how our same ideas can be applied with the use of her theory. Millikan points out that a representation is located between its producer, on the one hand, and the consumer of it, on the other hand. The producer of the representation must have as its function that it brings about a mapping between the representation and a state of affairs according to a set of mapping rules. For example, a bee can produce a bee dance such that the angle of the dance maps onto the location of nectar. Thus far the account is very similar to Dretske's; Millikan, however, emphasises that there must be a consumer of the representation, as well as a producer. The consumer uses the representation to carry out various functions, functions that have arisen out of an evolutionary and/or learning history. In the case of the bee dance, the consumers are other bees who use the dance to fly to the right location. For the bee dance, a certain single use is always made of the dance, but, in general, a representation that indicates a certain state of affairs can be put to all sorts of uses. If I represent "the chair is big", I can sit on it if I desire a big chair for sitting on, walk around it, ignore it, burn it, etc. But these uses can only be successfully carried out, to the extent that they are, if the state of affairs indicated by the representation actually holds. In fact, this is what defines the content of the representation, according to Millikan (1993, p. 89): The content of a representation is specified by the normal conditions for proper functioning of the consumers as they react to the representation. "Normal conditions" are those specified in a "normal explanation"; i.e. an explanation for how a proper function was successfully carried out in just those cases historically in which it was successfully carried out (thereby explaining historically why the device performing the function was selected).

A representation can be put to use by consumers partly because it exhibits a dimension or dimensions of possible variance parallel to possible sources of variance in the environment, as pointed out by Millikan. For example, in the bee dance, the angle of the dance varies in parallel with variation in the direction of the nectar. The intensity of the dance varies in parallel with the distance of the nectar. So the dance maps onto direction and distance; but these features of the environment do not exhaust the content of the representation. Millikan distinguishes between variant and invariant aspects of the representation (Millikan, 1984, p. 99). In our normal explanation of how the bee dance has been used historically (in just those cases in which it was successfully used), we must consider any feature that if removed from the environment or incorrectly mapped, will guarantee failure for its users (features that selection worked on to create or maintain as features that figure in the normal explanation). Thus the content of a bee dance "concerns the location of nectar, not [e.g.] the direction of the dancer's approach" to the hive (Millikan, 1993, p 109-110). The dance is about the location of nectar, but there is a difference between the contents "location" and "nectar": Location corresponds to a variant aspect of the representation (the representational medium varies with different locations), and nectar to an invariant aspect (nothing in the medium varies with the representation being about nectar or not nectar). It seemed to us (Dienes & Perner, 1999, 2001a,b) that this fundamental difference could be usefully understood in terms of an implicit/explicit distinction: The representation makes explicit that it is about location by having varying states for varying locations; it is not explicitly about nectar, because there is nothing in the representation that varies according to whether nectar is or is not the individual that has the represented location. We say that location has been explicitly represented in the bee

dance, but the fact that the representation is about nectar has been left implicit.¹ In general, distinctions are explicitly represented where variation in the representational medium corresponds to distinctions in the represented; the implicit content is the content of the representation that has not been explicitly represented.

This notion of implicit/explicit is just an extension of the everyday use of the implicit-explicit distinction in a specific way. When I say "The present king of France is bald" I have stated explicitly that the present king of France is bald, because this is the content of the representation, and, further, variations in the representational medium (words) would correspond to variations in this precise content. Moreover, we would say, in everyday terms, that the sentence implies (technically: presupposes) that there is a present king of France, but there is nothing in the medium that varies specifically with whether there is a current king of France or not. So it is natural to say that whether there is a present king of France is not represented explicitly, it is just conveyed implicitly.

Jimenez and Cleeremans (1999) wondered if our notion of "implicit" would allow implicit knowledge to have causal powers: If there is nothing in the representation that varies with nectar, how can the fact that it is about nectar have consequences for the system? Content, implicit or explicit, is defined in terms of normal conditions. Normal conditions are those that must be stated in a normal explanation of how the representation performs its function; i.e. they play a role in a causal story. The implicit content of a representation, just like the explicit content, must figure in a normal explanation and therefore must be causally efficacious (cf. Perner & Dienes, 1999). However, the fact that the dance is about nectar cannot be used for further inferences by the system of bees themselves, a point we return to shortly.

In the bee dance, the representation, in its role as an indicative representation, predicates certain properties (i.e. direction and distance) of the nectar. The properties are explicit, but the fact that they have been predicated of an individual (the particular supply of nectar) has been left implicit. We call this type of representation "predication implicit".

Millikan (1984) regarded bee dances as "intentional icons", as opposed to the fully-fledged representations that occur in many human mental states. Fully fledged representations differ from icons in that representations allow inference; specifically, they allow inference by identifying common elements across different representations as being the same. We equate this step in the first instance with predication explicitness, allowing the same individual to be tracked across different representations about that individual. Predication explicit representations have a subject-predicate structure, just what Millikan requires of fully fledged representations. The representation now not only makes explicit the property it is about, but also the individual that has the property and the fact that the one is predicated of the other. The full propositional content is thereby made explicit.

We argued that explicit inference in the sense of entertaining hypotheticals requires not only predication explicitness but also that the factuality of the proposition can be explicitly represented; e.g. whether the proposition is merely hypothetical, or is regarded as a fact. Factuality must also be represented to explicitly distinguish goals from reality, intentions and desires from beliefs, currently true from not now true but true in the past, and reality from counterfactual states (Perner, 1991; Evans & Over, 1999). The ability to represent factuality explicitly is thus an important step for a representational system. It is also needed for appreciation of phenomenal feel: For an organism to know that an experience is like anything, it must know the experience is similar or different to other

experiences, so it must know that the experience could have been otherwise (Perner, 2000). Thus, one can see the first link from explicitness to consciousness, a relation we will be dwelling on later.

So far we have described a hierarchy of ways in which a representation can be explicit; in this hierarchy, successively properties, a whole proposition, and factuality of the proposition are made explicit. There is one important final step. Dienes & Perner (1999) applied the notion of a hierarchy of explicitness specifically to what it is to have knowledge. When I know a fact, there is a person "I", who has an attitude of knowing towards a fact (a proposition with a certain factuality). For example, if I know by seeing a word in front of me that its meaning is "butter", the fact is "the word in front of me has the meaning butter". I could just represent the property "butter" (predication implicit representation). I could make explicit the full proposition "the word in front of me has the meaning butter". I could make explicit that the fact is indeed a fact, "it is a fact that the word in front of me has the meaning butter". Finally, I could represent explicitly that it is I who sees this fact, i.e. "I see the fact that the word in front of me has the meaning butter". This makes all aspects of the knowledge explicit. We call this final step in the hierarchy "attitude explicitness": One makes explicit the propositional attitude, or mental state, by which the fact is beheld (in this case, by seeing).

Implicit knowledge can thus be implicit in a number of ways, and we can already indicate some relations between the different levels of implicitness and the topics explored in this volume under the heading of the unity of consciousness. If seeing the word "butter" under degraded conditions allows only the predication-implicit representation "butter" to be formed, this representation, although maximally implicit, could be causally efficacious

in inducing me to say "butter" as the first dairy product that comes to mind, or in completing the stem "but---". However, I could not keep track of what individuals (objects, etc) have different properties (e.g. which of two words had the meaning butter). Creating a coherent world in the sense that single objects have bound to them the right properties requires predication explicit representations. That is, a representational system capable of predication explicit representations, the second level of explicitness, requires it to solve the binding problem, a central theme of this volume. However, even though properties have been bound to individuals, this would not necessarily enable a person to make a judgement about the facts of the bindings. In a factuality implicit representation, the representation is taken as true, but it has not been judged as being a fact rather than not a fact. For example, Bridgeman & Huemer (1998; see also Bridgeman, 1999; Perner & Dienes, 1999) found that people could track properties of each of two presented objects, as shown by their reaching behaviour, but they need not have represented the factuality of the facts, as shown by their poor judgements about the same facts (people could move their finger towards the true location of one of two objects, both subject to illusory motion, even while people reported incorrect locations of the objects; i.e. they reported the locations expected on the basis of the illusory motion). To be able to make a judgement as an act of judgement requires the factuality of the proposition be represented. In the Bridgeman and Huemer study, the representations guiding reaching were plausibly predication explicit (properties bound to individuals), but factuality implicit. When factuality has been left implicit, whether two propositions are contradictory or consistent is not explicitly represented as such (e.g. subjects were not aware of any discrepancy between their reaching and their verbal descriptions of events). However, appreciating the

"unity of consciousness" (or lack of) requires that the consistency of information can be represented. This role of factual explicitness in the unity of consciousness will be taken up later. First we will consider in more detail the step of making one's attitude of knowing explicit. What is gained at this step? The next stage in the argument will be to relate attitude explicitness to consciousness, and then to the unity of consciousness.

3. Attitude explicitness and consciousness

We will use the higher order thought theory to relate attitude explicitness to consciousness (Rosenthal, 1986, 2000, Carruthers, 1992, 2000; see Rosenthal, this volume). Rosenthal develops an account of when a mental state is a conscious mental state. He argues that when one is in a conscious mental state one is conscious of that mental state. It would be inconceivable to claim that one is in a conscious state of, for example, seeing the word butter, while at the same time denying being conscious of seeing the word butter. So the question is, how does one become conscious of mental states? I can be conscious of things in two ways. I can be conscious of you being there by perceiving you being there (e.g. by seeing you) or by thinking of you being there. We do not perceive mental states by any special sense organ that we know of; rather, Rosenthal argues, we think about them. We are conscious of our anxiety when we think that we are in an anxious state; we become conscious of our pain when we think that we are in pain; we become conscious of our seeing the word butter when we think that we are seeing the word butter. That is, when we are consciously seeing the word butter, we have a thought like "I see that the word is butter". Because this thought (this mental state) is about another mental state (seeing), it is called a higher order thought. In sum, the theory claims

that the necessary and sufficient conditions for having a conscious mental state is to have a higher order thought to the effect that one is in that mental state².

A fully attitude-explicit representation is exactly a higher order thought; it represents that one is in a certain mental state, e.g. seeing. Thus, full explicitness just is the necessary and sufficient condition for having conscious mental states, according to the higher order thought theory of Rosenthal. Fully explicit knowledge is conscious knowledge; conversely, knowledge that is not fully explicit is unconscious knowledge. But we have to be careful not to conclude that any representation that is not fully explicit is an unconscious mental state. We could not call the implicit representations produced by glucose detectors in the liver (that can be said to represent the glucose levels in the liver) unconscious mental states. The higher order thought theory is about mental states; it is only representations that correspond to mental states that can be unconscious mental states by virtue of being implicit.

Mental states are minimally states with content, i.e. they are always about something. One cannot think a thought without the thought being about something. But mental states typically have other properties as well; for example, bee dances are about nectar, but we do not regard such dances as mental states, nor do we regard states of glucose detectors as mental states although they are about something. Millikan (1984, 1993) points out that bee dances function as both indicative and imperative icons at the same time; they both indicate something in the world and have the function of bringing about a particular state of affairs in the world (flying to nectar, the imperative content of the icon). Mental representations, in contrast, often have these moods sharply distinguished (beliefs are just indicative, desires are just imperative). Mental

representations also sometimes appear to be such that that they can be concatenated while maintaining the same context-free meaning. But while true of some of our mental states, we do not logically have to require all our mental states to have these properties, so we cannot take these properties to be necessary properties of representations for the representations to be mental states. It may be true that at least some representations in the system as a whole need to have one or more of these properties (so that we could, e.g., attribute beliefs and desires to the system) for any representation in it to be regarded as mental. But these properties need not be possessed by each and every representation. For example, we are happy to call procedural knowledge "knowledge", but the indicative and imperative moods of procedural knowledge are not kept separate (see also Cotterill, this volume; Hurley, this volume). When we exhaust the conceptual description of our visual experience, the residual non-conceptual content is still part of our mental state of seeing despite not having a concatenative context-free meaning (Chrisley, 1996).

Without needing to specify the necessary and sufficient conditions for a state to be a mental state, we can specify some states that seem to us to be appropriately called mental without stretching the use of the term unacceptably. We will follow Carruthers (2000) in calling at least the following states mental states: Representations that are produced by our perceptual system and that guide the performance of actions correspond to mental states. For example, Carruthers argues that the states in the dorsal visual system controlling visuomotor performance have every right to be called visual mental states, precisely because they do guide our actions (even though they may never be available to be conscious). Similarly, enduring representational states in the brain caused by the perceptual system and that affect performance at an indefinite later time could, by the

same argument, be regarded as mental. For example, when a radiographer learns to interpret X-rays by many exposures changing the relative "weights" in him, we will consider the states of knowledge used in actively interpreting an X-ray as mental states, and thus as possibly being unconscious mental states. Similarly, we will regard the representational states of a person involved in producing and understanding language (these states having been triggered by appropriate perceptual input) to be mental. The reader need not agree with this stipulation, but it is exactly these sort of representational states that we will later argue can be shown to be implicit and unconscious.

4. Explicit knowledge and the unity of consciousness

We have just seen how fully explicit knowledge is conscious knowledge, by the higher order thought theory. The explicitness of knowledge is also what gives rise to various aspects of our sense of a unity of consciousness. Increases in explicitness are necessarily tied to greater unity, that is, they both make greater unity possible and require greater unity. Predication implicit properties are unrelated, predication explicitness makes it possible to represent their relationship, being part of the same or different object. Factuality helps relate factual and fictional propositions into a coherent view. The whole idea of factuality explicitness makes sense only with the notion of a world, which is just that which relates different propositions into a unified whole, i.e., every fact is related to other facts by having a place in the spatio-temporal space. Attitude explicitness allows relating different worlds into a coherent relationship: a possible world is related to the real world by being the thought world of a person in the real world. There is another

fundamental role played by attitude explicitness in explaining how and why consciousness seems, in some sense, unified to us. Attitude-explicit knowledge creates unity in two ways.

The first way arises because attitude-explicit representations involve representing mental states as related to an "I". Thus fully explicit knowledge which is related to the same "I" should be associated with a sense of being part of a unified consciousness (see Weisberg, this volume, for the development of this argument). Conversely, if different mental states were associated with different "I"s there may be a breakdown in the unity of consciousness. This is just what Kihlstrom (1997) suggests can happen in hypnosis. The highly hypnotizable person can create a separate "hypnotic I" capable of its own stream of consciousness not unified with the normal stream of consciousness. Thus, Hilgard (e.g. 1977, 1992) describes the "hidden observer". In his first experience with the hidden observer, Hilgard hypnotized a subject in class and suggested he was deaf. The subject then claimed not to hear anything of the conversation that ensued. At this point, it spontaneously occurred to Hilgard to suggest to the subject that there may be a hidden part of him that really did know what was going on and could be contacted whenever Hilgard touched his arm. Indeed, when Hilgard touched his arm, a hidden part (the hidden observer) responded and could recount the conversation the subject had previously denied hearing: It seemed there had been two simultaneous streams of consciousness. This phenomenon formed the basis of Hilgard's theory of hypnosis in general (hypnosis as divided consciousness). The interpretation of the hidden observer phenomenon is controversial (e.g. see Kihlstrom, 1998; Kirsch & Lynn, 1998; Woody & Sadler, 1998). But it is at least theoretically possible on the present framework if it is possible to "set up"

different "I"s voluntarily. What is involved in the latter and whether it is empirically possible requires a more developed theory of how we represent the I.

Why does representing a single I lead to an appreciation of a unity of consciousness?

Partly what constitutes what it is to be the same "I" is that states attributed to the "I" are mutually consistent. Consider the experiments in Marcel (1993). For one means of responding (e.g. verbally) the subject often asserted "I did not see the light flash", while at the same time with another means of responding (e.g. blinking, as a communicative gesture) the subject asserted "I did see the same flash". Marcel concluded there were different selves responding³. In sum, a single "I" requires a means by which states attributed to the "I" are made mutually consistent and can also be perceived as consistent. Thus, the unity of consciousness depends on higher order thoughts being consistent; consciousness awareness is produced by higher order thoughts, and inconsistent higher order thoughts therefore cause an incoherent conscious awareness. This issue of consistency brings us to the second way in which attitude explicit knowledge creates a sense of unity.

The second way by which attitude explicit knowledge creates a sense of unity is that it allows knowledge to be represented as knowledge and hence as coherent (or incoherent) with respect to other knowledge. Hence, explicit knowledge allows at least the perception of coherence. But it also does more than this.

All of us have numerous inconsistent beliefs. Most of the time we are unaware of these inconsistencies. Sometimes we are aware of inconsistency in our beliefs; it is at precisely such times that we feel a greater or lesser pressure to move towards more consistent beliefs. This means we are unlikely to believe consciously an object both has a

property and does not have a property; we are unlikely to believe consciously we are both in a mental state and that we are not in that mental state. That is, information that is simultaneously conscious tends to be made more consistent compared to information left unconscious, and this is part of what we mean by saying consciousness is unified. This unity arises because explicit factuality allows one to represent something as possibly true or false; it is not simply taken as true. In fact, we argue that one of the functions of explicit factuality is to explore possible implications of beliefs and coherence with other beliefs in order to create coherence.

One can imagine an automatic belief consistency checker that checked beliefs for consistency without ever representing factuality; it simply deletes one belief or the other. While this is logically possible, it is not optimal because there is no reason why the process should converge to a stable coherent set of beliefs. The inductive processes that produced the inconsistent beliefs are still left intact and capable of producing the same inconsistent beliefs again based on other existing beliefs. One cannot just randomly delete one of two existing beliefs that happen to contradict. What is needed is a mechanism that selectively deletes one belief or the other for good reason. Even better, it simply labels one belief as potentially to be acted on, and the other not to be acted on (functionally deleted, if you like). Labels are an advantage because such labels are in principle temporary, so the decision is defeasible. But if there is a normal explanation for the functioning of the labelling mechanism (i.e. that explains why it has been selected to act as it does) such that the labels work effectively because the label that allows the belief to be acted on corresponds to beliefs that are true and the label to not act on a belief corresponds to beliefs that are false, then the labels have the content "true" and "false" respectively; they

are explicit factuality markers. It is in fact highly likely that an effective labelling mechanism would have normal explanation just like this because true beliefs will tend to lead to appropriate actions and there is constant pressure on the system to create true beliefs anyway. That is, the mechanism would have been selected because of those cases (however scarce or common) in which it led to true rather than false beliefs being acted on⁴.

In any system which had the power to combine beliefs inferentially - in any system that was inferentially promiscuous to some extent - there would be pressure to evolve explicit factuality, for the reasons just given (it optimizes the generation of true beliefs). Perner (2000) argued that explicit factuality may be seen as a form of conscious awareness. One can see it in this way: the content of the factuality marker in the above scenario can mean, ambiguously, "is a fact" or "is-knowledge-for-me".⁵ That is, fact explicitness can have the status of a non-conceptual higher-order thought (cf Hurley, 1998). While this does not produce conscious awareness in the complete sense we can experience, it is part way there. Presumably a system capable of inference is along the road of conceptualizing what previously was non-conceptual knowledge in general. Therefore, once it has formed factuality markers, it is moving along the road to forming conceptual higher order thoughts - not necessarily for any particular selective reasons to do with higher order thoughts, but to do with whatever the general pressures were that made it conceptualize in the first place.⁶ One can thus see why inferential promiscuity is associated with consciousness; it is not sufficient for consciousness, but empirically one would expect it to be associated with conscious awareness. In sum, inferential promiscuity combined with factuality explicitness is part of the process by which the unity of consciousness is

determined and appreciated: conscious beliefs can be checked for consistency and made consistent.

Factuality-implicit knowledge (like the knowledge in a connectionist network) cannot be brought under the scope of the mechanism for explicitly detecting and producing consistency. This creates the possibility of simultaneous occurrent states of mutually incompatible implicit and explicit knowledge to co-occur indefinitely without being made consistent and without the subject being aware of an inconsistency⁷. An example of simultaneous active but contradictory implicit and explicit knowledge is provided by Bridgeman, Kirsh, and Sperling (1981), in which subjects point to a different location of an object than the one they verbally report. Another example is provided by Reed, McLeod, and Dienes (2001). When catching a cricket ball, we know that for a ball thrown towards a person they run forwards or backwards at speed that ensures the angle of gaze increases in a controlled way (Dienes & McLeod 1993; McLeod & Dienes, 1993). People report something else; they frequently report that the angle of gaze goes up and then, midway through the flight, goes down; or that it goes up and remains constant. They believe this even after being told to observe their angle of gaze during a successful catch. Thus, the two contradictory beliefs, the implicit and the explicit belief about angle of gaze, are simultaneously active but their inconsistency remains undetected.

If one held a theory that active mental representations (even if just constituting first-order mental states) are always conscious (e.g. Dulany, this volume; Perruchet, this volume), and higher order thoughts are only necessary for a certain type of consciousness (e.g. reflective consciousness, Block, 1995), then implicit knowledge is conscious knowledge that fails to be fully unified with the rest of consciousness; the investigation of

implicit knowledge is the investigation of the disunity of consciousness (Cleeremans, this volume). However, that is not our view; implicit knowledge is unconscious knowledge and therefore fails to take part in the mechanisms of creating unity that explicit knowledge takes part in. (Implicit knowledge will be driven towards consistency with other knowledge by any process that leads to true beliefs in explicit or implicit knowledge, e.g. learning rules appropriate for the environment the system evolved in; it's just implicit knowledge will not benefit from the extra mechanisms explicit consistency checking allows, and any consistency implicit knowledge has with other knowledge cannot be appreciated *as such*.)

5. Implicit knowledge in artificial grammar learning.

Our development of the nature of explicitness (and its links with the unity of consciousness) is empirically useful to the extent it can be translated into experimental practice. How does one go about determining in practice whether some knowledge is, for example, attitude implicit? In this section we will present novel evidence for unconscious implicit knowledge in people learning an artificial grammar to show how the ideas can be turned into experimental detail.

Reber (1967) introduced the artificial grammar learning paradigm and coined the term implicit learning to describe the process by which people attending to highly structured stimuli can acquire knowledge of the structure without being able to say what the structure is. A prototypical example is natural language; we soak up the rules but cannot say what they are. Not even linguists have an explicit statement of the complete grammar of a natural language, yet we have learnt it implicitly. Reber wondered if he could observe the process in the laboratory by studying how people learn artificial

grammars. Chomsky had just specified a hierarchy of grammars and it was natural to start with the bottom of the hierarchy, the finite-state grammar. For thirty-five years since then, we have all followed suit. This will be the first published implicit learning study we know of that had a look at the next level up the hierarchy: the context-free grammar. However, for the purposes of this paper it is not important what people learnt; merely that they learnt something. We just note as a point of side interest, to be developed as a main point in a different publication, that the content of the experimenter's grammar is qualitatively different from previous grammars used in the literature (for a review, see Dienes & Berry, 1997).

The point of the experiment for current purposes is to illustrate how our framework can be used to explore whether knowledge acquired in a domain is unconscious or not, an issue that has dogged researchers for years. We argue that our framework provides a firm foundation for taking metacognitive measures -assessing the extent to which people know that they know, for example - as an appropriate tool for establishing the conscious or unconscious status of knowledge. We first have to establish that the subject is in certain mental states, different states of knowing (guessing, reasonably sure, etc). Then we see if the subject has represented that they are in these different states by taking confidence ratings. Have they formed relevant higher order thoughts (attitude explicit representations) that they are in certain mental states? If they have not, the knowledge is attitude implicit and unconscious. Implicit learning is just the learning mechanism that produces such attitude implicit knowledge. It is just such knowledge that fails to take part in the unifying mechanisms operating on explicit knowledge.

Previous research has investigated our methodology with finite state grammars (see Dienes, Altmann, Kwan, & Goode, 1995; Dienes & Berry, 1997 for a review). In this experiment we used a particular type of context-free grammar one could call the A^nB^n rule. There is a set of elements called the A set and another set called the B set. To make a grammatical string, start with any number, n, of A's in any order and follow by the same number of B's in any order. In our experiment, the elements were letters and the A set was the letters M, V, and Q; the B set was R, X, and T. So, for example, for n=3, MQMXXT is grammatical. In addition, we added randomly one or two letters from a different set (comprising the letters A and Y) to the start of the string, to disguise the rule and allow grammatical items to be both of odd and even length.

In the training phase, subjects copied down 40 strings. They were not told there were any rules involved, for all they knew the strings were just random sequences of letters. In fact, 20 of the strings were n=2 grammatical strings and 20 of the strings were n=4 grammatical strings. In the test phase, subjects were informed that actually the set of strings they just saw obeyed a complex rule, and they were asked to classify 60 strings as obeying the rule or not. In fact, 20 of the strings were n=2 strings; 10 of these were made nongrammatical by adding or deleting an A letter or a B letter; 20 were n=3, with 10 of these made nongrammatical; and 20 were n=4, with 10 of these made nongrammatical. The test strings had one further constraint: all the bigrams in them were novel, in the sense that they did not appear in the training phase. A bigram is a sequence of two consecutive letters, e.g. QM. Because all the bigrams were novel, all the n-grams in general were novel. That is, subjects could not perform the task by learning which bigrams (in general,

n-grams) had appeared in the training phase, so prima facie the task could not be learnt by, for example, the general purpose learning rule suggested by Perruchet (this volume)⁸.

There were two groups of subjects. The 10 control subjects just received the test phase without any training; the 10 trained subjects received both the training phase and the test phase. Figure 1 below shows the percentage correct classification of the test strings by the trained and control subjects (note: the standard error for any one bar in the figure was not more than 4.3).

Insert figure 1 about here

As can be seen, the trained subjects classified at a greater level than the control subjects, $F(1,18) = 9.19$, $p = .007$. There was no interaction of training with n-value, $F < 1$: Not only did the trained subjects learn to classify novel strings of the same n value they were trained on ($n=2,4$), but they could also interpolate to an n of 3. But the point for us is simply that there was learning.

Subjects were in fact tested twice on each string twice. So we can look at the proportion of times a subject got a string correct twice (CC), in error and then correct (EC), correct and then in error (CE), or in error twice (EE) (cf Reber, 1989), i.e. the subjects pattern of consistency in their responding. Figure 2 displays the results.

Insert Figure 2 about here

Figure 2 shows that for trained subjects:

(a) There is a substantial amount of inconsistent responding (i.e. EC and CE); that is, subjects did not just follow some deterministic rule. Instead they responded to a string with a certain probability of saying "grammatical". Reber (1967, 1989) suggested that subjects may respond with three different probabilities: 1 or 0.5 to grammatical strings and 0 or 0.5 to nongrammatical strings. On this model, one would expect the frequency of EE to be the same as EC or CE; indeed, there was no significant difference between EE and the average of CE and EC, $t(9) = 1.17$, $p = .27$ in these data. Similarly CC should be greater than the average of EC and CE, which it was, $t(9) = 2.81$, $p = .020$. We argued (Dienes, Kurz, Bernhaupt, & Perner, 1997) that (with the more standard finite-state grammars) different subjects used a wider range of different probabilities of responding grammatical to each item; this more detailed analysis has yet not been performed with the type of grammar used in this experiment. But the distinction between Dienes et al (1997) and Reber is not important for the main point: Subjects did not respond deterministically, and the data are consistent with subjects using a range of probabilities for responding "grammatical" to different strings.

(b) Consistent responding is associated with making the correct response (CC vs EE, $F(1,9) = 15.13$, $p = .004$).

(a) is evidence that the learning/knowledge application system is treating itself as having different degrees of knowledge about different situations (it responds "grammatical" with a range of probabilities to different strings); (b) is evidence that it got this correct. That is, the subject is in fact in different knowledge states, not just from our perspective, but from the subject's as well. But have the subjects conceptualized

themselves as being in the different knowledge states that they are in fact in? Dienes and Perner (1996,1999) argued that the knowledge is implicit if this information about degree of knowledge is not transmitted through the subjects' confidence ratings.

After every classification decision subjects gave a confidence rating on a 50% to 100% scale. They were informed that 50% meant literal guessing, they really had NO information about the right decision and they could just as well have flipped a coin. 100% meant complete certainty. Figure 3 displays the relationship between subjects' accuracy and their confidence.

Insert Figure 3 about here

When subjects said they were literally guessing, they were in fact performing significantly above chance with a classification performance of 65% (SD = 20%), $t(9) = 2.31$, $p < .05$. That is, subjects did not know that they knew. Further the slope of the regression line was non-significantly different from zero, $F < 1$. That is, subjects did not know when they were in different knowledge states. On both these grounds, the knowledge is attitude implicit.

In sum, people can consistently treat some knowledge of an artificial grammar as knowledge without knowing they have knowledge: The knowledge is self and attitude implicit. Such knowledge does not form part of a unified consciousness either because (a) it is unconscious (our view); or (b) it is conscious but not unified (the possibility raised by Cleeremans, this volume).

Footnotes

1. Independently of our development of these concepts, Chandler (2000) used a very similar notion of explicit and tacit belief to deal with a quite different set of issues.
2. There is in addition a further stipulation specified by Rosenthal: The higher order thought must not be the result of any inference of which we are conscious. In order for the mental state to be a conscious mental state, we must not consciously infer that we are in that mental state.
3. In fact, in this case, we do not interpret Marcel's results as indicating different selves, a "slippage in consciousness", but rather as indicating the unconscious slipping in different responses on those measures that are less under conscious control, ie., that are less declarative. The results showed that the less declarative the response (eye blink vs verbal response), the more accurate the response, as predicted by us but not by the different selves hypothesis.
4. Once factuality markers had evolved for the stated reason, it became logically possible to the system to act on representations labelled as non-factual in an act of pretence.
5. The factuality marker's meaning is ambiguous in this way if the representational system used does not explicitly distinguish "fact" from "knowledge for me". If the representational system does make the distinction, then factuality explicitness does not constitute a non-conceptual higher order thought and hence is not sufficient for a form of conscious awareness. Thus, in adult humans, fact explicitness is arguably not sufficient for conscious awareness.

6. Once conceptual higher order thoughts became possible, occurrent representations of factuality plausibly in general led to the formation of the appropriate higher order thought as an immediate inference (albeit not explicitly represented as an inference); Gordon (1995) calls this an ascent routine. The ascent is not guaranteed to happen; hypnosis provides a case in point of failed ascent (Dienes & Perner, 2001b).
7. Two explicit beliefs can be simultaneously conscious without their inconsistency being detected. But in this case there are mechanisms (in particular, tagging for factuality) that have the function of detecting consistency even if they fail to do so. These same mechanisms are not operative when implicit knowledge is involved.
8. In other experiments we have ruled out that knowledge of abstract repetition structure (Brooks & Vokey, 1991) or which position which letters occur in accounts for the degree of learning observed.

References

- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18, 227-287.
- Bridgeman, B. (1999). Implicit and explicit representations of visual space. *Behavioral and Brain Sciences*, 22, 759-760.
- Bridgeman, B., & Huemer, V. (1998). A spatially oriented decision does not induce consciousness in a motor task. *Consciousness and Cognition*, 7, 454 - 464.
- Bridgeman, B., Kirch, M., & Sperling, A. (1981). Segregation of cognitive and motor aspects of visual function using induced motion. *Perception and Psychophysics*, 29, 336-342.
- Brooks, L. & Vokey, J. (1991). Abstract analogies and abstracted grammars: Comments on Reber (1989) and Mathews et al (1989). *Journal of Experimental Psychology: General*, 120, 316-323.
- Carruthers, P. (1992). Consciousness and concepts. *Proceedings of the Aristotelian Society, Supplementary Vol. LXVI*, 42-59.
- Carruthers, P. (2000). *Phenomenal consciousness naturally*. Cambridge: Cambridge University Press.
- Chandler, J. (2000). On being inexplicit. Cognitive Science Research Papers 516, School of Computing and Cognitive Sciences, University of Sussex.
- Chrisley, R. L. (1996). Non-conceptual psychological explanation: Content and computation. D.Phil thesis, University of Oxford.
- Dienes, Z., Altmann, G., Kwan, L., Goode, A. (1995) Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 21, 1322-1338.
- Dienes, Z., & Berry, D. (1997). Implicit learning: below the subjective threshold. *Psychonomic Bulletin and Review*, 4, 3-23.
- Dienes, Z., Kurz, A., Bernhaupt, R., & Perner, J. (1997). Application of implicit knowledge: Deterministic or probabilistic? *Psychologica Belgica*, 37, 89-112.
- Dienes, Z., & McLeod, P. (1993) How to catch a cricket ball. *Perception*, 22, 1427-1439.
- Dienes, Z., & Perner, J. (1996) Implicit knowledge in people and connectionist networks. In G. Underwood (Ed), *Implicit cognition* (pp 227-256), Oxford University Press.
- Dienes, Z., & Perner, J. (1999) A theory of implicit and explicit knowledge. *Behavioural and Brain Sciences*, 22, 735-755.
- Dienes, Z., & Perner, J. (in press) A theory of the implicit nature of implicit learning. In Cleeremans, A., & French, R. (Eds), *Implicit learning and knowledge of the real world*. Psychology Press.
- Dienes, Z., & Perner, J. (forthcoming). The metacognitive implications of the implicit-explicit distinction. In Chambres, P., Marescaux, P.-J., Izaute, M. (Eds), *Metacognition: Process, function, and use*. Kluwer.
- Dretske, F. (1988). *Explaining behaviour: Reasons in a world of causes*. Cambridge (Massachusetts), London: The MIT Press.
- Evans, J. B. T., & Over, D. (1999). Explicit representations in hypothetical thinking. *Behavioral and Brain Sciences*, 22, 763-764.
- Gordon, R. M. (1995). Simulation without introspection or inference from me to you. In M. Davies & T. Stone (Eds), *Mental simulation: Evaluations and applications*. Blackwell.
- Hilgard, E. R. (1977). *Divided consciousness: Multiple controls in human thought and action*. New York: Wiley.
- Hilgard, E. R. (1992). Dissociation and theories of hypnosis. In E. Fromm, & M. R. Nash (Eds), *Contemporary hypnosis research* (pp. 69-101). London: The Guilford Press.
- Hurley, S. L. (1998) *Consciousness in Action*. Cambridge: Harvard University Press.

- Jimenez, L., & Cleeremans, A. (1999). Fishing with the wrong nets: How the implicit slips through the representational theory of mind. *Behavioral and Brain Sciences*, 22, 771.
- Kihlstrom, J. F. (1997). Consciousness and me-ness. In J. D. Cohen & J. W. Schooler (Eds), *Scientific approaches to consciousness*. New Jersey: Erlbaum.
- Kihlstrom, J. F. (1998). Dissociations and dissociation theory in hypnosis: Comment on Kirsch and Lynn (1998). *Psychological Bulletin*, 123, 186-191.
- Kirsch, I., & Lynn, S. J. (1998). Dissociation theories of hypnosis. *Psychological Bulletin*, 123, 100-115.
- Marcel, A. J. (1993). Slippage in the unity of consciousness. In G. R. Bock and J. Marsh (Eds), *Experimental and theoretical studies of consciousness: the Ciba Foundation symposium 174*. New York: Wiley.
- McLeod, P., & Dienes, Z. (1993). Running to catch the ball. *Nature*, 362, 23.
- Millikan, R. G. (1984). Language, thought, and other biological categories. Cambridge, MA: MIT Press.
- Millikan, R. G. (1993). *White queen psychology and other essays for Alice*. Cambridge, MA: Bradford Books/MIT-Press.
- Perner, J. (1991). *Understanding the representational mind..* Cambridge, MA: MIT Press. A Bradford Book.
- Perner, J. (2000). Dual control and the causal theory of action. In N.Eilan & J.Roessler (Eds.). *Agency and self-awareness*. Oxford: Oxford University Press.
- Perner, J., & Dienes, Z. (1999). Deconstructing RTK: How to Explicate a Theory of Implicit Knowledge. *Behavioural and Brain Sciences*, 22, 790-808.
- Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behaviour*, 6, 855-863.
- Reed, N., McLeod, P., & Dienes, Z. (submitted). Implicit knowledge and motor skill: What people who know how to catch a ball don't know.
- Rosenthal, D.M. (1986). Two concepts of consciousness. *Philosophical Studies*, 49, 329-359.
- Rosenthal, D.M. (2000). Consciousness, Content, and Metacognitive Judgments, *Consciousness and Cognition*, 9, 203-214.
- Woody, E., & Sadler, P. (1998). On reintegrating dissociated theories: Comment on Kirsch and Lynn (1998). *Psychological Bulletin*, 123, 192-197.

Figure 1

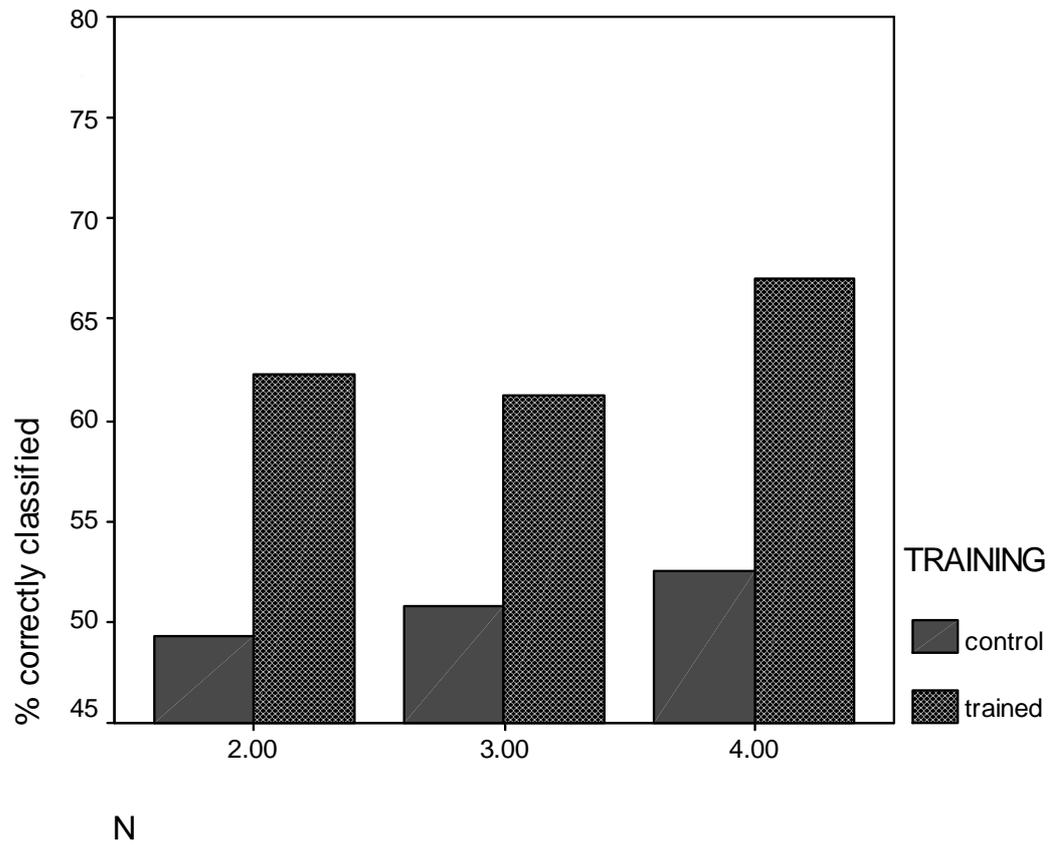


Figure 2

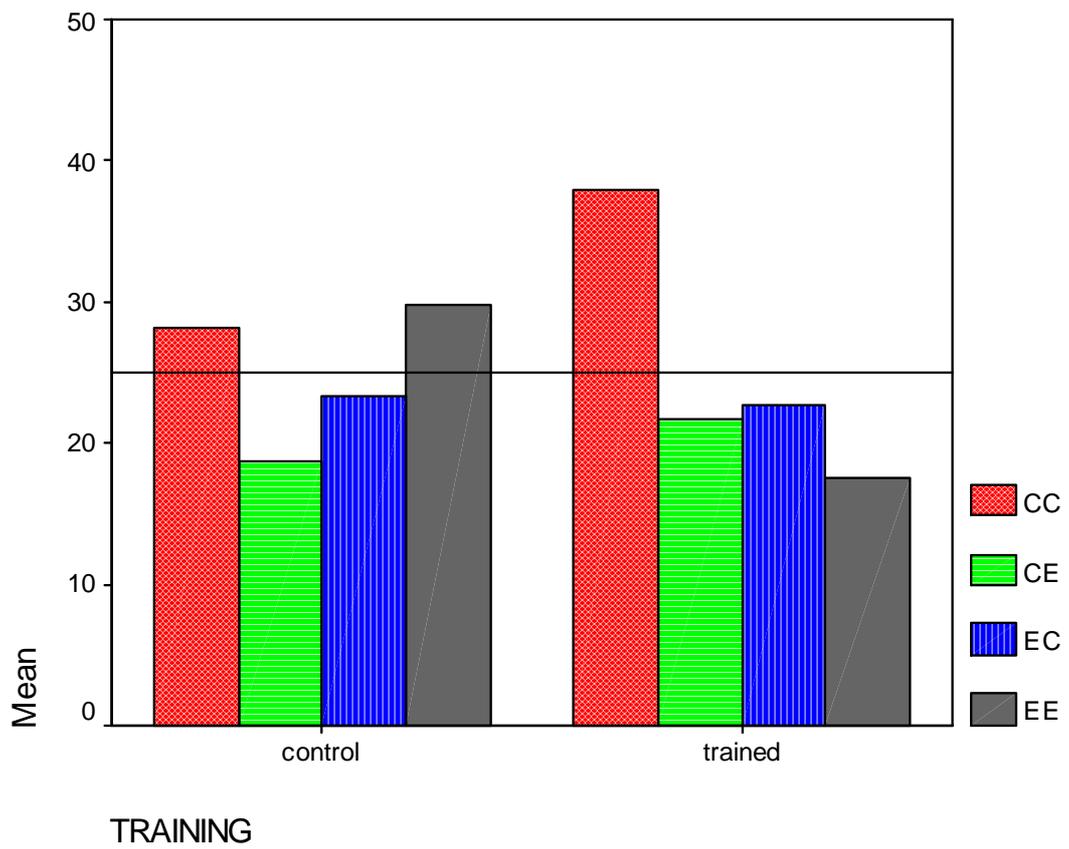


Figure 3

