A theory of implicit and explicit knowledge

Zoltan Dienes
Experimental Psychology
University of Sussex
Brighton
Sussex BN1 9QG
England
dienes@epunix.susx.ac.uk
http://www.biols.susx.ac.uk/faculty/ep/dienes.htm

Josef Perner
Institut fuer Psychologie
Universitaet Salzburg
Hellbrunnerstrasse 34
A-5020 Salzburg
Austria
josef.perner@sbg.ac.at
http://www.sbg.ac.at/psy/people/perner_e.htm

Dec, 1998

KEYWORDS:
Implicit knowledge, consciousness, automaticity, memory, cognitive development, visual perception, artificial grammar learning

Acknowledgements.

We wish to thank Bruce Bridgeman, John Campbell, Peter Carruthers, Martin Davies, Ron Chrisley, R.Carlson, Greg Currie, Tony Marcel, Shawn Nichols, Gabriel Segal for invaluable discussions and Peter Carruthers, John Kihlstrom, Pierre Perruchet, and Carol Seger for their informative reviews.

Short Abstract.

The ordinary meaning of information being implicit or explicit is applied to the different aspects of knowledge, resulting in a partial hierarchy of ways in which something can be known implicitly or explicitly. The most important type of implicit knowledge consists of representations that merely reflect the property of objects or events without predicating them of any particular entity or event. The clearest cases of explicit knowledge are reflective representations that represent one's own attitude of knowing. The relationship to similar existing distinctions (procedural-declarative, conscious-unconscious, verbalizable-non verbalizable, direct-indirect tests, automatic-voluntary control) and potential applications to research areas concerned with the implicit-explicit distinction (visual perception, memory, cognitive development, and artificial grammar learning) are discussed.

Long abstract

The implicit-explicit distinction is applied to knowledge representations. Knowledge is taken to be an attitude towards a proposition which is true. The proposition itself predicates a property to some entity. A number of ways in which knowledge can be implicit or explicit emerge. If a higher aspect is known explicitly then each lower one must also be known explicitly. This partial hierarchy reduces the number of ways in which knowledge can be explicit. In the most important type of implicit knowledge representations merely reflect the property of objects or events without predicating them of any particular entity. The clearest cases of explicit knowledge of a fact are representations of one's own attitude of knowing that fact. These distinctions are discussed in their relationship to similar distinctions such as procedural-declarative, conscious-unconscious, verbalizable-nonverbalizable, direct-indirect tests, and automatic-voluntary control. This is followed by an outline of how these distinctions can be used to integrate and relate the often divergent uses of the implicit-explicit distinction in different research areas. We illustrate this for visual perception, memory, cognitive development, and artificial grammar learning.

Objectives.

The objective of this target article is to provide an analysis of the distinction between implicit and explicit knowledge in terms of the semantic and functional properties of mental representation. In particular this analysis attempts to:

(1) create a common terminology for systematically relating the somewhat different uses of the implicit-explicit distinction in different research areas, in particular, learning, memory, visual perception, and cognitive development;
(2) clarify and generate predictions about the nature of implicit knowledge in different domains;
(3) clarify why the distinction has traditionally been brought into close contact with notions such as consciousness, verbalizability, voluntariness-automaticity, etc.;
(4) justify why different empirical criteria (e.g., subjective threshold, objective threshold, direct-indirect tests) are used to identify implicit/explicit knowledge;
(5) justify the use of the implicit-explicit terminology by observing the ordinary language meaning of ”implicit” and ”explicit”.

Our basic strategy for meeting these objectives is to analyse knowledge as a propositional attitude according to the representational theory of mind (RTM; Field, 1978; Fodor, 1978). Roughly speaking, if I know a fact (e.g., the animal in front of me is a cat) then, according to RTM, I have a representation of that fact and the internal, functional use of this representation constitutes it as knowledge of mine (rather than a desire of mine, etc.). Knowledge can vary depending on what is represented (made explicit) and which aspects remain implicit in the functional use of representations. This application of the implicit-explicit distinction has several advantages.

The main advantage of our analysis is that it provides a common ground for the use of the implicit-explicit distinction in different fields of investigation. Consider Schacter’s (1987) influential definition of the implicit-explicit memory distinction: ”Implicit memory is revealed when previous experiences facilitate performance on a task that does not require conscious or intentional recollection of those experiences; explicit memory is revealed when performance on a task requires conscious recollection of previous experiences.” This definition may capture the phenomenal experience of implicit and explicit memory very well, but it leaves open how the definition is to apply to implicit and explicit knowledge in other fields. For example, Karmiloff-Smith (1986, 1992) has argued that there are several steps of ”explicitation” before consciousness is reached. Identifying being explicit with being conscious gives us no understanding of why Karmiloff-Smith’s lower forms of explicitness have anything to do with this distinction. In other words, although it has been suggested that the implicit-explicit dichotomy should be broken into a series of explicitness levels our analysis is needed to explain just what it is that becomes more explicit as one ascends levels and to relate levels in one research area to different subdivisions of explicitness in other areas.

Existing problems of this kind with the implicit-explicit distinction are many. In research on memory and subliminal perception, explicitness has been linked to performance on direct versus indirect tests (Richardson-Klavehn & Bjork, 1988; Reingold and Merikle, 1993) because direct

test performance seems to require conscious awareness. The interesting question left open, however, is why direct tests require consciousness. Or, in visual perception, it is found that touching an object is based on unconscious, implicit information whereas pointing to the object requires conscious, explicit information that is subject to visual illusions (e.g., Bridgeman, 1991; Milner & Goodale, 1995, Rossetti, 1997). Why? More directly, what are the representational requirements for conscious awareness? What is the relation between knowledge over which we have voluntary control and knowledge of which we are aware? Why can we sometimes control in limited ways knowledge of which we are not aware (Dienes, Altmann, Kwan, & Goode, 1995)? Can predictions be made for the conditions under which knowledge will be represented implicitly? With our analysis of the implicit-explicit distinction, we are able to give some answers to these questions.

Another advantage of our analysis is that it is grounded in the ordinary use of the terms ”implicit” and ”explicit” (e.g.: ”They didn’t say so explicitly, it was left implicit”), whereas traditional definitions have depended on further related distinctions. Schacter (1987, p. 501) defined implicit memory by its lack of conscious or intentional recollection, and Reber (1993, p. 5) defined implicit learning as ”the acquisition of knowledge that takes place largely independently of conscious attempts to learn and largely in the absence of explicit knowledge about what was acquired.” These definitions of implicit memory/learning raise the question of why the terms implicit/explicit are used at all. Why not call explicit memory or learning directly by their name, that is, conscious memory or conscious learning (cf Reingold and Merikle, 1993, p. 42)? Moreover, when using technical terms with an existing ordinary meaning, it seems to us, we should adhere to that existing meaning as far as possible and not impose some arbitrary ‘operational definition’, or else we make it difficult for the scientific community to share the same meaning, because the natural meaning is likely to keep intruding. (Who still adheres—or ever adhered—to the operational definition of intelligence as that which the WAIS measures?). So it is not an unimportant feature of our use of the implicit-explicit distinction that it attempts to stay true to its natural meaning, which we believe was the unarticulated reason for introducing the distinction in the first place, and what partially motivated its acceptance and continued use.

We ordinarily say that a fact is conveyed explicitly if it is expressed by the standard meaning of the words used. If something is conveyed but not explicitly, then we say it has been conveyed implicitly. We can discern two main sources of implicitness. One is the contextual function/use of what has been said explicitly. A prime case is *presuppositions*. To use a famous example, the statement, ”The present king of France is bald,” presupposes that there is a present king of France. It does not express this fact explicitly because the function of the sentence (when uttered as an assertion) is to differentiate the present king of France being bald from his not being bald. For that reason the speaker of this sentence can claim that he <u>did not</u> (explicitly) say that there was a king of France. Yet the presupposition does commit him to there being a king of France, or else his assertion of the king being bald becomes insincere. So in this sense he did (and thus we say: ”implicitly”) convey that there is a king of France.

The other source of implicitness lies in the conceptual structure of the explicitly used words. For example, if one conveys that a person is a *bachelor*, then one conveys that this person is *male* and *unmarried* without making those features explicit. Using ”bachelor” commits oneself quite strongly to ”male” and ”unmarried” lest one shows oneself ignorant of the meaning of the word

bachelor in the language spoken. These are not rare cases. Whenever we say that something is an X (e.g., a bird), we implicitly convey that it is also an instance of the super-ordinate category of X (e.g., an animal) on the same grounds as in the bachelor case.

It is common to both sources of implicitness that the information conveyed implicitly concerns *supporting facts* that are *necessary* for the explicit part to have the meaning it has. The implicitly conveyed fact that *there is a king of France* is necessary for the explicitly expressed information that *he is bald* to have its normal, sincere meaning. Similarly, that someone is male and unmarried is a necessary supporting fact for the explicitly conveyed fact that he is a bachelor.

In our analysis the distinction is between which parts of the knowledge are explicitly represented and which parts are implicit in either the functional role or the conceptual structure of the explicit representations. A fact is explicitly represented if there is an expression (mental or otherwise) whose meaning is just that fact; in other words, if there is an internal state whose function is to indicate that fact.[1] Supporting facts that are not explicitly represented but must hold for the explicitly known fact to be known are *implicitly represented*.

2. The Representational Theory of Knowledge.

2.1 Implicitness arising from functional role.

Mental concepts such as knowledge are standardly analysed as propositional attitudes (Russell, 1919). The sentence ”I know that this is a cat” consists of a person (I), a proposition (this is a cat) and an attitude relation between person and proposition (knowing). The representational theory of mind (Field, 1978; Fodor, 1978) is concerned with how such an attitude can be implemented in our mind. The suggestion is that the proposition is represented and the attitude results from how that representation is used by the person (functional role). The representation ”this is a cat” constitutes knowledge if it is put in what philosophers would call *”knowledge box”* or cognitive scientists would call a *data base*. The representation is used as a reflection of the state of the world and not as it would be, for example, if it were in a *goal* box, as a typically nonexisting but desirable state of the world.

In this view we can say that the content of the knowledge is explicit because it is represented by the relevant representational distinctions (by analogy with explicit verbal communication). That is, there is an internal state whose function is to indicate the content of the knowledge. In contrast, the fact that this content functions as knowledge is left implicit in its functional role[2] (as implicitly conveyed information is communicated by the functional necessities created by the explicit part). The fact that it is I myself who hold this knowledge is not explicitly represented, it is implicit in the fact that I do hold that knowledge. We accordingly have three main types of

1 This requires that there be a system that can go into at least two states, one state for the fact and another either for the negation of the fact or for staying noncommittal about the fact.

2 There is no provision in this system for being in one state to indicate this is knowledge and being in another state either to leave it open whether this is knowledge or or to indicate that it is not knowledge.

explicit knowledge, depending on which of the (3) constituents of the propositional attitude is represented explicitly:

(1) explicit content but implicit attitude and implicit holder (self) of the attitude.
(2) explicit content and attitude but implicit holder of attitude.
(3) explicit content, attitude and self.

This large picture has to be refined in at least three ways. First, the same shift from implicit to explicit also applies within each constituent, complicating the picture somewhat. Second, arguments are needed as to why only the above combinations occur and not all the other logically possible ones (e.g., an explicit representation of self but implicit attitude and content). We start by discussing the refinements required for the first type of each of the three constituents of propositional attitudes.

2.1.1 Content

The content of a propositional attitude, like knowledge, is what the attitude is about. In our example of the cat that I see in front of me, I know that it is a cat. The representation of the content of this knowledge as ”this is a cat” identifies (1) a *particular individual* (i.e., the animal in front of me), (2) a *property* (or natural kind: catness), and (3) it *predicates* this property of the particular individual. For a more succinct and more general way of expressing these aspects we use predicate calculus notation, where F, G,... denote properties, a, b, ... denote particular individuals, and the syntactic combination of F and b into the formula Fb expresses that F is predicated of b.

Even though this content makes these three elements explicit, however, there are other aspects that remain implicit. For example, I clearly know that the individual is now a cat, and that it is a fact about the real world that it is a cat, not just a cat in some fictional context. That is, (4a) the temporal context of the known state of affairs and (4b) its factuality are left implicit.

We have identified (4) main components of a known fact about which we can ask whether they need to be represented explicitly or can be left implicit:

(1) properties, e.g.: ‘F’, ‘being a cat’.
(2) individuals, e.g.: ‘b’, ‘particular individual in front of me’.
(3) the predication of the property to the individual, e.g.: ‘Fb’, ‘this is a cat’.
(4) temporal context and factuality (vs. fiction),
e.g.: ‘It is a fact of this world that at time t, Fb’, ‘It is a fact that this is currently a cat’.

The question is now whether any of these components can remain implicit and whether they can remain implicit independently of each other or only in certain combinations. We argue that they can only remain implicit in roughly the order in which they are listed above, i.e., if an element with a higher number is represented explicitly then every element of a lower number must also be represented explicitly.

As an extreme case in which almost everything is left implicit we consider Strawson’s (1959, p. 206) ”naming game”, in which a person simply calls out the name of a presented object, e.g., ”cat” or ”dog”, depending on which kind of animal is presented. In this context, the word

”cat” expresses knowledge of the fact that ‘this (object in front of the person) is a cat’ and conveys this information to the initiated listener. We could not say anything less, for example, that it only expresses knowledge of cat-ness, or of the concept of cat. Yet, what are made explicit within the vocabulary of this naming game are only the properties of being-a-cat, being-a-dog, etc. Consequently, since there is knowledge that it is the particular presented individual that is a cat or dog, that knowledge remains implicit.[3]

Our use of Strawson’s naming game only provides an example of the property (cat) being represented explicitly, the individual and predicating the property of this individual remain implicit. The naming game uses the publicly inspectable medium of language, but, when it comes to the question of which aspects can be made explicit independently of other aspects, it becomes an imperfect guide for explicitness of mental representations, as the following shows.

In the naming game, it is also possible to represent individuals explicitly and to leave their properties implicit. This is the case for forced choices between two items, by pointing to the item that has a particular property, for example, which one of two objects - the left or the right - is a cat. In the case of the naming game, one could argue that the response must explicitly distinguish the two items (a, b) by pointing right or left, but not the property. The pointing thus conveys the information ‘This one is a cat’ but makes only ‘this one’ explicit and leaves ‘is a cat’ implicit. In the case of the naming game (i.e., the information passing between two communicating parties) this is possible. In the case of the knowledge that a single person must bring to bear, explicitness of the individuals requires explicitness of the attributed property, because the person must be able to go into a cat/no-cat state for each individual in order to decide which is a cat and then to respond correctly. Hence, for knowledge we have the constraint that explicit representation of the individual to which a property is attributed entails explicit representation of that property.

At this point one should be made aware that the notion of predicating something of a particular individual need not be restricted to particular objects or persons. It will be applied later in extended form to events and even to causal regularities. Traditional logic does not make this very explicit but Barwise and Perry’s (1983) Situation Semantics offers an elaborate distinction between event types and individual events, in order to capture the capacity of natural language to freely reference particular events, causal regularities, laws, etc. and then to describe them as having certain properties or as being of a certain type. For example, a particular event (b) was a dance (F) and has the further feature of having had me as a participant (G) etc.

Subliminal perception provides an example from psychological research, as discussed in more detail in Section 3.2. The suggestion is that under subliminal conditions only the properties of a stimulus (the kind of stimulus) get explicitly represented (e.g., the word ”butter”), not the fact that there is a particular stimulus event that is of that kind. This would be enough to influence indirect tests, in which no reference is made to the stimulus event (e.g., naming milk products), by raising the likelihood of responding with the subliminally presented stimulus (”butter” is listed as a milk product more often than without subliminal presentation). The stimulus word is not

---

3 As a point of interest, one should mention that what remain implicit in this case are *unarticulated* constituents of what is known (Perry, 1986) in the sense that they do not find expression in the representational vehicle. As a result, the knowledge remains "situated" within the causal context of knowledge formation and inferences drawn from this knowledge are valid only as long as this context is maintained (Barwise, 1987; Fodor, 1987).

given as response to a direct test (e.g., Which word did I just flash?) because there is no representation of any word having been flashed. Performance on a direct test can be improved with instructions to guess (Marcel, 1993), because this gives leave to treat the direct test like an indirect test, just saying what comes to mind first.

As mentioned earlier, even explicit representation of F being predicated of b (”Fb”, or ”This is a cat”) leaves implicit the fact that Fb is a true proposition, i.e., a fact at the present time. Only the representation ”Fb is a fact now” represents *the fact that b is F at the present time* completely explicitly. The reason for making these aspects explicit may seem superfluous. In particular, the addition ”is a fact” may strike some readers as totally redundant and trivial, so let us dwell briefly on its significance.

Consider a simple mental system that does not represent truth explicitly but just contains a single model of how it perceives the world to be (Perner, 1991, described the young infant as having only this representational power). The model of the world is a type of knowledge box in that any proposition Fb that is in the knowledge box is taken (judged) to be true, on the grounds of being in that box plus the functional role the box plays in the mental economy. There is no possibility of representing propositions that are not true, without creating mental havoc, however, because all propositions in the box are acted upon as if they were true (Leslie, 1987, pointed this out in his analysis of pretence). To differentiate true from false propositions, one could represent false propositions in a different functional box, as has been suggested for pretence and for counterfactual reasoning (Currie & Ravenscroft, in press; Nichols and Stich, 1998). In concrete terms this means that a child who is pretending that the banana is a telephone represents, "this is a banana (Bb)" in its knowledge box and, "this is a telephone (Tb)" in its pretend box. This solution may be adequate for pretend play consisting of switching from a knowledge (serious action) mode into a pretend mode of functioning. Pretend actions are then simply governed by the representations inside the pretend box. This cannot account for the child knowing what it is pretending. To know that, the pretend representations have to be in the knowledge box. This raises the problem of cognitive confusion (representational abuse, Leslie, 1987) and the pretend representations have to be quarantined in some sort of "metarepresentational[4] context" (Sperber, 1997). Such markers explicitly differentiate within the knowledge box what is to be taken as true from what is not to be taken as true. More generally, for knowing what is and what is not truethe truth value has to be made explicit within the knowledge box, that is, to represent ”Fb is a fact” or ”Fb is NOT a fact”.[5] This distinction is also required for understanding change over time (i.e., to represent that Fb was the case and now Gb is the case; Perner, 1991; 1995, Appendix) and to interpret symbolic expressions and representations (e.g., to understand that objects in the world are also in the picture).[6]

---

4 "Metapresentational" here is used in the looser sense of modifying representational status (as used by Leslie, 1987) and not in its usual strong meaning of representing the representational relationship (Pylyshyn, 1978) as Perner (1991) has pointed out.

5 Representation of the truth of Fb does not replace the functional role of the knowledge box of mentally asserting Fb, a problem Frege grappled with in his "Begriffsschrift" (see Currie, 1982, ch. 4). But it allows representation of false propositions within one's knowledge box without their becoming asserted. That is, by representing "Fb is not a fact", in the functional role of knowledge, Fb is represented but not asserted. What is asserted is that Fb is not a fact.

6 Perner (1991) reviews evidence that these abilities – pretend play, understanding temporal change and understanding representations – emerge at about the same age of 18 months.

The following table gives a summary of the different cases of the possible implicit-explicit combinations of facts that we have discussed so far. We suggest that these are the only realistically possible ones.

Insert Table 1 about here

Table 1 excludes certain permutations of the four components property, individual, predication and factuality. For the verbal exclamations in Strawson's naming game all combinations are possible, but for knowledge only the four cases listed above are possible. For example, predication cannot be known explicitly on its own. It can be explicitly conveyed on its own in the naming game in response to the question "Does b have the property F?" The response "Does/doesn't have it" represents only predication explicitly. Again, a system that can do this must make further internal distinctions, it must distinguish F from not-F in order to decide whether the presented object "does/doesn't have" that property. Knowledge of the presented individual can remain implicit. This case is accounted for in 2(b) above.

In the case of factuality we are after the distinction between whether a state of affairs Fb is a fact or fiction. The naming game can only be played with real objects. A system that can meaningfully distinguish between whether the predication of F to b holds in the real world or in a world of fiction must have the representational resources to specify the property and the individual in question and be able to predicate this property of the individual in order to decide whether it holds in reality or only in fiction. Hence, if factuality is known explicitly then predication, individual, and property must also be known explicitly. Similarly, the time of a fact can only be left implicit for the present. A system that can distinguish between whether the predication of F of b holds now or in the past must have the representational resources to specify the property and the individual in question and be able to predicate of this property of the individual in order to decide whether it holds now or has held previously. Hence, if time is known explicitly then predication, individual, and property must also be known explicitly.

Memory research provides a relevant example. Explicit memory is not only conscious, but, more to the point, a recollection of the past. For this it must represent past events as having taken place in the past. Only then can systematic answers be given to direct questions about the past. If a past event is only represented by its properties (event structure) then it can influence indirect and direct tests alike. Only when the pastness of the event is represented explicitly can performance on a direct test that addresses the pastness directly outshine performance on indirect tests (see Reingold and Merikle's, 1993, criterion for explicit memory). So we can see why and how test directness relates to explicitness. In the next section we see how it relates to consciousness.

2.1.2 Attitude

Knowledge is standardly analysed as propositional attitudes. The system knows some fact (e.g., the fact that b is F,or that this is a cat) if it is related in a particular way to the proposition expressing that fact. In the representational theory of mind this is the case if the following conditions hold:

(o) The system has a representation, R, of this fact, and
(i) R is accurate (true),

(ii) R is used by the system as an accurate reflection of reality (i.e., the system must *judge* that b being an F is the case), and
(iii) R has been properly caused (it must not have come about by accident, but (?) have a respectable causal *origin*, which when made explicit serves to *justify* the claim to knowledge).

Possession, *accuracy*, *judgement* and *causal origin (justification)* are all supporting facts for any representation to constitute knowledge. E.g., ”Fb is a fact” constitutes knowledge of *the fact that b is F* for a system only if (o) the system has the representation, (i) it is accurate, (ii) it is treated by the system as an accurate reflection of the world (the world is judged to be so) and (iii) it came about in a proper causal (justifiable) way. Hence all four facts are implicit in any knowledge until made explicit.

These four facts define the *attitude* of knowledge. Making them explicit means making the attitude explicit. For that the system has to form the following metarepresentations, where R stands for the representation of the known fact (i.e., R = ”Fb is a fact”):

(0) ”R is possessed by the system”
(1) ”R accurately reflects the fact that Fb.”
(2) ”R is being taken (judged) as accurately reflecting the fact that Fb.”
(3) ”R was properly caused by its content through a generally reliable process (i.e., it is caused by the fact Fb through the reliable process of visual perception).”

In other words, (0) represents that the knowledge content can be entertained by the system, (1) represents the knowledge as a true thought (that is, as a true thought that is being *merely entertained* but *not judged* as being true, see Künne, 1995), (2) represents the knowledge as a belief, and (3) represents the knowledge as causally justified thought.
Only if the system can entertain R as a representation that it possesses can it represent what further properties (e.g. (1), (2), and (3)) this representation might have. But the three further metarepresentations can be explicit independently of each other. Truth does not imply having been properly caused nor being taken for true; being taken for true does not imply either being true or being properly caused; and having been properly caused does not imply being taken for true or being necessarily true because, although generally reliable, even such a process can on occasion fail[7]. Note that some dependencies emerge if one represents that it is the same rational agent (e.g. oneself) who represents R as accurate and who represents R as being taken to be true.

If (0)-(3) hold, then the system represents its *attitude of knowing* explicitly, i.e.:
”There is knowledge of the fact that Fb”. What this does not make explicit is the holder of this attitude, i.e., the self. The fact that it is oneself who holds the attitude is implicit in the act of knowing. To make it explicit, the system has to represent itself as the holder of the attitude: ”I know that Fb is a fact”.[8,9]

---

7 We are grateful to Peter Carruthers for having pointed out in response to an earlier draft that without this addition ”through a generally reliable process” our criterion (3) and with it our definition of knowledge becomes otiose. The practical point of criterion (3) is to distinguish reliable from unreliable sources, but even the most reliable source can in principle fail. If one requires that process to be so reliable that it necessarily follows that it produces true representations then criterion (3) would imply criterion (1), but at the cost of a practically useless criterion (3).

8 This self-explicitness can be applied separately to the (4) different aspects of knowledge:
(0s) ”I possess R”
(1s) ”I have R which accurately reflects the fact that Fb.”
(2s) ”I take (judge) R as accurately reflecting the fact that Fb.”

Other attitudes may be held towards a piece of knowledge, e.g. I *guess* that Fb is a fact. Making any attitude explicit always requires (0) to hold, plus additional representations, depending on the attitude.

2.1.3 Relating Explicitness of Content, Attitude and Self.

It is evident that explicit representation of self as holder of an attitude (e.g., "I know ...") contains an explicit representation of the attitude ("know"). The interesting question concerns the degree to which explicit representation of knowing requires explicit representation of the content (e.g., this is a cat). That is: Is it possible to explicitly represent ”I know” or ”it is known” and leave implicit the fact that *this is a cat (Fb)*. In a variation of the naming game an expression like, "I know," can be implicitly conveying that the knowledge is of the fact that Fb. However, inside a (rational) agent this explicit reflection on knowledge implies explicit factuality of the known; one must be able to judge the factuality of the known fact before coming to the conclusion that one knows that fact. Since explicit factuality implies explicitness of predication, individuals, and properties, we can conclude that explicit representation of self or attitude implies explicit representation of the content.

The dependencies we have discussed are summarised in Figure 1. If an aspect at a higher level is represented explicitly (at the origin of an arrow) then—according to our analysis—all aspects at a lower level (at the end of the arrow) must also be explicitly represented.

Insert Figure 1 about here

On the basis of this partial hierarchy we will later speak conveniently of knowledge that is "fully explicit" when all aspects are explicitly represented, "attitude-explicit" when everything up to the attitude is explicit, and "content-explicit" if all the aspects of content are represented explicitly. "Attitude-implicit" will indicate that attitude and all higher aspects in the hierarchy are left implicit, and so on for the other aspects. It is also often convenient to differentiate between different levels within content: "fact-explicit" (equivalent to "content-explicit") when all aspects of content are explicit, "predication-explicit" when predication, individuals and property are made explicit (for simplicity’s sake we ignore the possibility of case 2b in Table 1), and of "completely implicit" if only properties remain explicit.

(3s) ”I have R which has been properly caused by its content through a generally reliable process, e.g., I saw the fact Fb”.

The following implications hold between these three types of self-explicitness for a rational agent who takes himself to be rational: (1s), (2s), and (3s) each imply (0s). (2s) implies (1s) because representing oneself as believing Fb implies that one represents Fb as true. In other words, one cannot represent oneself as believing something that one represents as false. Conversely, (1s) implies (2s) because if one represents R as true one should treat it as true. (3s) strongly suggests but does not strictly imply (1s) (and hence (2s)), since representing that the knowledge was properly caused implies that it ought to be accurate (i.e., that I should take it to be accurate).

9 Conditions (0), (i), (ii), and (iii) capture the everyday use of the word ‘know’. Cognitive scientists generally use a broader definition namely, requiring only conditions (0), (ii) and (iii) to hold; simply being false is not sufficient reason to prevent a piece of knowledge from being knowledge (e.g Newton’s Laws). Removing conditions (i) and (1) would not alter any of the conclusions that follow; note that (1s) given in footnote 6 should still be included, as it follows from (2s), so our characterization of fully explicit knowledge stands as is.

On an important cautionary note, one must point out that these hierarchical constraints only hold for a single representation. That is, a single representation cannot make something explicit at the higher level and still represent aspects at a lower level implicitly. This does not preclude the possibility of two independent representations, one of which making something explicit at the higher level and the other representing something at the lower level implicitly. For instance:

(a) ”I know that there is some fact involving F”
(i.e., explicitly representing attitude and factuality).

(b) ”F” (i.e., implicitly representing predication of F to b).

This is possible, but the point is that (a) does not implicitly represent the fact that Fb. Rather, it explicitly represents the knowledge that there is something concerning the property F. In that case, there is no implicit knowledge of Fb being a fact. That this is not implicit in (a) can be seen from the fact that Fb is not a supporting fact of (a), i.e., one can know that there was something about F without the fact that Fb.

2.2 Implicitness Due to Conceptual Structure.

This kind of implicitness (*structure implicitness*) typically arises when the system represents (has a concept for) properties that can be defined as compounds of more basic properties; for example, the property of being a bachelor has the components of being male and unmarried. So if one explicitly states that a person is a bachelor, then one implicitly conveys that he is also unmarried, because being unmarried is a necessary, supportive fact for being a bachelor. Similarly, one can explicitly know that someone is a bachelor, but not explicitly know that he is not married. However, as not being married is a necessary fact for being a bachelor, this fact is known implicitly. In this example, the structure of the component properties (male, unmarried, etc.) remains implicit in the explicit representation of the compound property (being a bachelor): a case of ”property-structure implicitness”. Roberts and MacLeod (1995) argued that concepts acquired incidentally and nonstrategically may have nondecomposable atomic representations in which the property structure is represented implicitly in our terminology.

2.3 Summary.

We have so far developed a rich structure for describing different ways some knowledge can be implicit within the use of some other explicitly represented knowledge. That is, knowledge with explicit representations of part of its content can contain other parts of its content: the attitude and self as holder of the attitude implicitly. Also, explicit knowledge can be a representation of compounds (typically: compound properties) that leaves the structure of its components implicit. We now explore how our analysis unifies the different distinctions that have traditionally been used in connection with the implicit-explicit distinction.

3. Related distinctions and test criteria.

The previous section showed that knowledge can differ in how many of its functional and conceptual aspects that are represented explicitly. This puts us into a position to show that the various distinctions associated with the implicit-explicit distinction differ in the amount of explicit representation required. We start with consciousness, as it has been used most prominently to define explicit knowledge (in memory, Schacter, 1987; in learning of rules, Reber,

1989). We will show that under a common understanding of "conscious", knowledge counts as conscious only if its content, the attitude of knowing, and the holder of that attitude (self) can be represented explicitly. Hence, conscious knowledge is, indeed, prototypically explicit.

Consciousness has often been related to (even defined in terms of) verbalizability (e.g., Dennett, 1978). The ability to address the content of one's knowledge verbally (direct tests) has often been used to test conscious and explicit knowledge. This makes sense in our analysis, because verbal reference requires very explicit representation of content. Furthermore, a close relative of verbally expressible knowledge, "declarative" knowledge, has often been put in opposition to "procedural knowledge." Although this opposition confounds several independent dimensions (procedural-inert, declarative-nondeclarative, and accessible-inaccessible), we can explain why these groupings appear natural and why they can be tied to the implicit-explicit distinction. Finally, the ability to exert voluntary control, in contrast to automatic action, has been associated with explicit, conscious knowledge. This link is justified, because, voluntary control requires explicit representation of one's attitude, which conforms to the requirement for conscious awareness, whereas automatic action can be sustained by procedural know-how.

3.1 Consciousness

We use "consciousness" (some philosophers might find the term "conscious awareness" more appropriate[10]) here as (we think) most people use it; one's knowledge is available to oneself and it is not necessary to prove its existence to one's own surprise through behavioural evidence. This is certainly the meaning of the conscious-unconscious distinction in cognitive psychology, as we will see from the many research examples in the next section. For example, implicit unconscious memory occurs where I appear to have no knowledge (memory) of a past event but can be shown by behavioural evidence in an indirect test to have some (implicit) knowledge of that event.

The idea that consciousness has something to do with the awareness of our mental states has a venerable tradition dating back to at least the writings of John Locke (cit. Tye, 1995, p. 5): "consciousness is the perception of what passes in a Man's own mind" and perhaps even to Aristotle (Güzeldere, 1995, p. 335). This intuition has recently been given prominence under the name of the Higher-Order-Thought Theory of Consciousness. Different versions of this theory differ as to the nature of the second-order state required. Armstrong (1980), like Locke, sees it as a perceptual state, a higher order act of observing our first order mental states--, Rosenthal (1986) sees it as a more cognitive state, and Carruthers (1996) as a potential for being recursively embedded in higher-order states (see Güzeldere, 1995). The basic insight behind these different approaches is that when one is conscious of some state of affairs (e.g., that the banana in my hand is yellow) then one is also aware of the mental state by which one beholds that state of affairs (i.e., I <u>see</u> that the banana is yellow). There is something intuitively correct about this claim, because it is inconceivable that I could sincerely claim, "I am conscious of this banana being yellow" and at the same time deny having any knowledge of whether I see the banana, or hear

10For example Dretske (1995) speaks of being "conscious" or "aware" when we have information about something and represent it as such as shown by the appropriateness of our behaviour. In this usage what we have in mind needs to be expressed as being "consciously aware" to distinguish it from being "unconsciously aware" (which some might find a strange combination, because "aware" or "conscious" carries the connotation of being consciously aware).

about it, or just know of it, or whether it is I who see it, etc. That is, it is a necessary condition for consciousness of a fact X that I entertain a higher mental state (second order thought) that represents the first order mental state with the content X.

Of course, there is philosophical controversy about whether this characterisation can capture the whole phenomenon of consciousness or just an aspect of it.[11] We need only focus on the less controversial part of this theory, namely, that the higher order mental state is merely necessary, although, in what follows we will occasionally explore the explanatory power of the stronger theory (that a higher order thought is both necessary and sufficient for consciousness). To be safe we will pursue Carruther's "potentialist" version of the higher order thought theory in more detail. Because it does not require actualty entertaining a higher order thought but only the potential for forming such a higher order thought, it makes less demands on the cognitive complexity of routine conscious information processing than the other versions of this theory. This potentialist version is sufficient for explaining why consciousness relates to explicitness, verbal expressibility, voluntary control, etc.

Carruthers (1996) sees consciousness as the potential of our mental content to be recursively embedded in higher order states. In other words, the content X of a knowledge state is conscious if it is recursively accessible to higher order thoughts, e.g., knowing that I know that X. In order to form this second order state one must to explicitly represent the first order knowing. For this, one in turn needs to represent the content explicitly, in particular its factuality, i.e., "it is a *fact* that X". This is a necessary condition. Tt is not always necessary to have the first order attitude and the self explicitly represented because these can be freely inferred from the factuality of the content as Gordon (1995) has pointed out in the context of simulation theory. Within one's own perspective--and that is all we are concerned with here--there is a one to one correspondence between what is a fact for me and what I know. Gordon speaks of ascent routines that allow us to go from descriptions of facts to knowledge attributions for oneself, e.g., from "X is a fact" I can go to "I know that X". That means that once factuality is represented explicitly, explicit representation of attitude and self is also possible. Of course, other conditions may have to be met (e.g., the representation must be in a short term memory store), but explicit representation of factuality (and thus all other aspects of content) is often all that is required.

In sum, on the weak version of the higher-order thought theory where potential access for higher order thoughts is only a necessary condition, we can conclude that the explicit representation of self and attitude is necessary for conscious knowledge but sometimes only the explicit representation of factuality is all that is needed. On the stronger version, where access for higher order thoughts is also a sufficient condition, explicit representation of self and attitude or factuality is sufficient for conscious knowledge. For us the critical implication of this view of consciousness is that the requisite higher order states represent the attitude and the holder of the first order state explicitly. This in turn requires explicit representation of the content of the first order mental state. This means that to have conscious knowledge one must represent all three aspects of it explicitly (or be able to form such explicit representations). For example, to

---

11 Block (e.g., 1994, 1995) emphasises the subjective feel of conscious experiences (phenomenal consciousness) as central to the mystery of consciousness. Our concern and that of most cognitive sciences would be merely a case of "access consciousness" or "monitoring consciousness". There are, however, some interesting arguments to the effect that second order mental states are necessary and sufficient for subjective feel (e.g., Carruthers,1992, 1996).

consciously know that the banana is yellow, I must explicitly represent that it is a present fact that the banana is yellow, that this fact is known and I must be able to explicitly represent that it is me who knows it. This analysis makes clear why most definitions of explicit knowledge involve consciousness; because it imposes the clearest, most extreme case of explicitness. It also puts us in good position to understand why verbal access to knowledge (and other features to be discussed below) are tied to consciousness.

3.2 Verbalisation and directness of tests.

In this subsection we wish to show why verbal access to knowledge is considered a sign of explicit, conscious knowledge, relating this to the important types of direct and indirect tests and the objective and subjective thresholds of perception.

Verbal communication (for transmitting information) proceeds by predication. A referring expression (or an ostensive gesture) is used to identify an individual (topic) and then further information about this individual follows. Hence, verbal report requires knowledge with explicit predication. An even stronger explicitness is necessary for the following reason. Linguistic information unlike perceptual information cannot be taken at its face value. As Gibson (1950) has emphasised, visual perception is highly reliable under most normal circumstances and thus can—barring the few visual illusions—be taken as veridical. Applied to linguistic information, this strategy would lead to a highly unstable knowledge base (Perner, 1991, chapt. 4). For this reason, verbal information needs to be interpreted without being taken as true prima facie. Only after evaluation (checking compatibility with other available information) should it be accepted. A distinction must accordingly be made between 'is a fact' and 'not yet clear', i.e., factuality has to be represented explicitly.

In research on implicit memory (Richardson-Klavehn & Bjork, 1988) and subliminal perception (Reingold & Merikle, 1988), a critical distinction is made between direct and indirect tests of knowledge. A *direct test* is one that refers to the fact in question. An *indirect test* does not refer to the fact in question, but the answer to some unrelated question or the response to some stimulus shows that some information about the fact must still be present. In both literatures, the fact in question is the spatio-temporal context of the presentation of a particular stimulus. The key methodological difference between implicit memory and subliminal perception is how long after the presentation of the stimulus knowledge of this fact is tested (Kihlstrom, Barnhardt, & Tataryn, 1992). In implicit memory, the fact in question could be that a particular word was studied 10 minutes ago in the laboratory, and typically the word is consciously perceived at the time of study. Implicit memory is considered in more detail in section 5.2 below. In subliminal perception, the fact in question is whether a particular stimulus has *just* been presented. According to the normal approach (e.g. Holender, 1986), perception is subliminal or implicit (Kihlstrom et al, 1992) if the participant performs at chance on a direct test of some aspect of this fact (because it was not consciously perceived), but the stimulus still affects processing indirectly.

Our analysis makes clear why performance on direct and indirect tests has something to do with implicit-explicitness and consciousness of the probed knowledge, provided the test questions are ”bona fide”, with participants saying that X is the case only if they have a representation stating that X is a fact. The analysis also makes it clear, however, that one cannot

equate test performance with type of knowledge, since there is no guarantee that test answers are bona fide, participants might say that X is the case just on the basis of a feeling that that might be right.

Even knowledge without explicit predication can influence indirect test responses, because the test does not refer to the event in question. For example, after a brief (e.g., 10 msec) presentation of the word "doctor" or "table" followed (within, e.g., 50 msec) by a patterned mask (a technique for inducing subliminal perception), a clearly visible word (e.g., "nurse") or nonword (e.g., ”nurge”) is presented and observers must judge whether or not this item is a word; this lexical decision provides an indirect test of knowledge of the first word. Although the instructions refer only to the clearly visible word, it has been found (e.g., Marcel, 1983a) that the identification of "nurse" is faster if the first word is semantically related (i.e., "doctor") than if it is unrelated ("table"). For this processing advantage to occur it is sufficient to take in only the property of the presented stimulus, i.e., "doctor" without any representation that there was a particular event that had that property. The semantic processing triggered by the word form "doctor" will activate the semantic field of the medical profession which then gives "nurse" a greater processing advantage than "table".

In contrast, a direct test refers to the event in question. There are different ways of making this reference. The question can refer to the event, e.g. "What was the word on the screen?". A bona fide answer "doctor" can be given only if the event has registered *as a fact*. So we see that bona fide performance on such a direct test requires explicit representation of factuality which, on Carruthers’s higher-order theory of consciousness is at least a necessary and possibly also a sufficient condition for consciousness. This provides a theoretical justification for using direct tests to assess conscious knowledge if all answers are bona fide. Unfortunately, there is no guarantee of this. Co-operative participants in our experiments try to give the best answer, and then even knowledge with implicit predication (far removed from meeting the criterion for consciousness) may help them give correct answers (based on guesses) on direct tests, a known problem in the field (e.g., Roediger and McDermott, 1996).

Performance on indirect tests can be influenced by conscious knowledge as well as implicit knowledge lacking explicit predication. One could only infer the use of implicit knowledge without consciousness from the advantage in performance of an indirect over a direct test (even if non bona fide answers are given on the direct test). This conclusion is warranted especially if performance on the direct test outstrips performance on the indirect test under conscious processing conditions so that any lingering issues about sensitivity differences (Shanks & St John, 1994) are eliminated (Reingold and Merikle, 1993, p. 53).

Since direct tests do not typically involve reference to one's subjective mental state of seeing, Cheesman and Merikle (1984; see also Greenwald, 1992) referred to the threshold conforming to this test as the "objective threshold": If the interstimulus interval between a stimulus (e.g. a word) and a mask is reduced so as to make perception more difficult, the objective threshold is defined by the interstimulus interval at which the participant performs at chance on a direct test of the nature of the stimulus presented. Our analysis, however, suggests that this might not reflect a single threshold, because there are at least two significantly different ways of making such a reference (cf Dagenbach, Carr, & Wilhelmsen, 1989). One is to stipulate

that an event occurred with the observer's task being to determine of which type the event was, e.g.: "What was the word on the screen?" This way of questioning puts the focus of the observer's mental search on finding a suitable property for an answer. A predication implicit representation of the perceived property will serve that purpose.

A different way of phrasing the question is to stipulate a particular event type, for example, the occurrence of a word; the observer's task is to decide whether or not it took place (i.e., to judge the existence or occurrence of a word). Marcel's (1983a, Experiment 1) query about whether a word (any word) was *present* or *absent* to determine the detection threshold appears to be of this kind. Here observers had to judge whether or not a word occurred. Such a judgement would require a predication-explicit representation of the perceived event. A mere representation of the property 'word', without explicit predication of the observed event, would not provide a natural answer to the observer's mental search initiated by the presence-absence question. Several studies and replication attempts inspired by Marcel's work used the other approach for determining the detection threshold, for example, "Which colour word was it (one of four possible colours)?" (Cheeseman & Merikle, 1984) or "Was there a word or a blank?" (Dagenbach, et al., 1989). In this case, a predication implicit representation of the event type ("red" or "word" or "blank") provides an answer for the mental search. This may be one reason these studies had only partial success in replicating Marcel's original finding that detection (absence-presence) has a higher threshold (i.e. occurs at a longer stimulus onset asynchrony, SOA, between stimulus and mask) than graphic or semantic similarity judgements (also see Fowler, Wolford, Slade, & Tassinary, 1981).

There is also the possibility of formulating a direct test by referring to the target event as a perceptually experienced event: "What was the word that you just saw?". For a bona fide answer the stimulus event must be encoded explicitly as a *visually perceived event.* Without that encoding the observer can only answer "I didn't see anything".[12] Since reflection on one's state of seeing is required, this detection criterion corresponds to the "subjective threshold" introduced by Cheesman and Merikle (1984, 1986; see also Merikle, 1992); the point at which participants know they know what they saw.

This discussion was mainly intended to show that the known problems in this field can be formulated in our framework. The contamination of explicit (direct) tests by implicit knowledge and of implicit (indirect) tests by explicit knowledge has been debated particularly intensively in memory research. As a solution, Jacoby (1991) proposed his process dissociation procedure, which brings in voluntary conscious control as an arbiter. We will discuss the relation between the implicit-explicit distinction and consciousness and volition in the next two sections.

3.3 Procedural versus declarative knowledge and accessibility.

The notion of procedural and declarative knowledge has been related to the implicit-explicit distinction by several authors. Karmiloff-Smith (1986, 1992) characterized as procedural

12 This is exactly what a blindsighted person will say, when performing at random. The critical trick that Weiskrantz, Warrington, Sanders, and Marshall (1974) used to get more convincing performance than Pöppel, Held, and Frost (1973) did was to instruct the patient to guess: "I'll show you a light that you won't be able to see. Even though you can't see it, give it a guess and point to it." (Weiskrantz, 1988, p. 187)

implicit knowledge that is severely limited in its accessibility to other parts of the system. Accessibility has been emphasised as the central factor in the distinction between procedural and declarative knowledge by Kirsh (1991). Squire (e.g., 1992) characterized the knowledge of the past that is typically impaired in amnesics as declarative memory (where declarative is considered largely a terminological variant of explicit memory or ”knowing that”), he contrasted this with nondeclarative (implicit, knowing how) memory, which includes procedural memory (habits, skills and conditioned reactions) but also memory of facts revealed by priming.

Our own suggestion is that at least four different dimensions are in play and need to be kept conceptually distinct: knowledge that is or is not contained in a procedure, declarative vs. nondeclarative knowledge, accessibility, and implicitness vs. explicitness. The goal, however, is to show that there are some necessary relations between these dimensions and the types of knowledge that form natural clusters: procedural knowledge tends to be implicit and hence inaccessible, whereas declarative knowledge involves quite explicit representation of its content, and hence tends to be conscious and accessible for different uses.

To some, implicit knowledge may simply mean inaccessibility. Apart from being an arbitrary conceptual stipulation, this definition of implicitness also lacks precision. Inaccessible in what way? All knowledge must be accessible in some way or it would not qualify as knowledge (on views like those of Millikan, 1984; Dretske, 1988); in any case, there would be no evidence that there was any knowledge at all. Our framework indicates how the implicitness of different aspects of knowledge makes it inaccessible in different ways, as indicated in our discussion in section 3.2 on direct and indirect tests and verbalizability, and in our treatment of procedural knowledge, which we now discuss.

The procedural-declarative knowledge distinction was introduced in artificial intelligence (McCarthy & Hayes, 1969; Winograd, 1975) and later taken over in psychological modelling by Anderson (e.g., 1976). It concerned how best to implement knowledge: Should one represent the knowledge that all men are mortal as a general declaration "for every individual it is true that if that individual is human it is also mortal"? Whenever knowledge of a human individual was introduced in the data base this general information would be consulted to infer by general inference rules that that individual must also be mortal. The alternative would be to have a specialised inference procedure: "Whenever an individual is introduced that is human, represent that that individual is mortal." [13]

Now we can see in what sense declarative knowledge is explicit. It represents explicitly that the regularity 'if human then mortal' is predicated of individuals and its general application to every individual is also marked. Moreover (provided the data base provides the requisite expressive power), it states that this regularity is a fact. In contrast, the procedure that adds 'is

13More technically expressed, the issue was whether one should represent the knowledge that every man is mortal as (1) a declarative axiom "$\forall x$ (Human(x) $\supset$ Mortal(x))" and then apply the general inference procedure "[$\forall x$ (F(x) $\supset$ G(x)) and F(b)] G(b)" which means roughly: If in the data base you find for Variables F, G, x and b the expressions "$\forall x$ (F(x) $\supset$ G(x)" and "F(b)" then add "G(b)" to the data base, or (2) should one encode the relevant knowledge directly in a specialised procedure: "Human(b) Mortal(b)". Our interest is in the difference between representing the regularity that being human implies being mortal either by means of the declarative implication sign "$\supset$" or by means of an inference procedure (production) symbolised as " ".

mortal' to every human individual it encounters, also knows something about this regularity but its knowledge is implicit in its application; its generality is implicit in the fact that it is applied to every encountered individual. But there is no distinction made in the system that represents that it is applied to individuals and that it is applied to every individual. The analysis also brings out the intuitive meaning of declarative knowledge as knowledge that declares what is the case (e.g., Squire, 1992, p 204: memory whose content can be declared) because it represents explicitly that something is a fact. Nondeclarative memory can be given precision in our analysis either as the stronger form of knowledge that does not make predication explicit or as a weaker form of knowledge that makes predication explicit but leaves factuality implicit.

The implicit nature of procedural knowledge also makes it clear why it has limited accessibility. For example the implicitness of the procedural representation of the fact that all humans are mortal does not allow the distinction between whether this rule applies to a current case and my thinking about the rule. To separate these two cases one needs some internal distinction that (explicitly) represents whether or not the rule applies. Then one can distinguish whether one is just thinking about the rule without it actually applying, or whether one is thinking about it because it applies. Moreover, there is no way of checking the adequacy of procedural knowledge. Such a check requires explicit representation of factuality in order to represent the result of the inference as a hypothetical possibility for comparing it with other available evidence.[14] All this puts a severe limitation on the usability of procedural knowledge.

The advantage of procedural knowledge is its efficiency. Procedures need not search a large database because the knowledge is contained in the procedures. Knowledge that resides in the application of a procedure, leaves predication and factuality implicit. As a result, it is limited in its accessibility in a way that has been claimed for modularity (Fodor, 1983); modular knowledge, for example, applies only to a specific input modality, it cannot use knowledge from other domains, etc. Implicitness of procedural knowledge is accordingly a natural basis for modularity in our input modalities which do not require fact explicit representation. In this context modular knowledge can be called implicit. However, implicitness is a less natural ally of modularity in the case of central processes (Fodor, 1987, ”modularity gone mad").

Modular or quasi-modular central conceptual processes have been proposed by Cosmides (1989) for reasoning processes that use a cheating detector module. Sperber (1996) considers quasi-modularity a general feature of central cognition. Smith and Tsimpli (1995, ch. 5) posited a quasi-modular central language module to explain isolated highly developed foreign language ability in an otherwise handicapped individual. The central language module is not the same as the usual linguistic input processing module, because it is not used to converse in different languages, but to translate playfully from one language into another. Such central modules are unlikely to operate purely procedurally without explicit predication or factuality. This is very clear in Leslie’s (1987, 1994) theory-of-mind module, proposal to explain the relative ease and speed with which children develop a theory of mind. A theory of mind does not just process factual information. It must represent the content of people's beliefs and desires, hence explicit

14 It might appear that learning systems, which are based on purely procedural knowledge, can make this evaluation on the grounds of negative feedback. The critical difference is that negative feedback in learning leads to a weakening of the response tendency for future inferences but it leaves the already made inference uncontested.

representation of factuality is required. Modular knowledge in this sense clearly cannot be implicit in the sense defined in this paper. [15]

In sum, knowledge contained in the application of a procedure (procedural knowledge) is active and efficient, but it leaves predication and factuality implicit: hence it is nondeclarative and limited in its range of applicability (hypothetical reasoning, checking validity) and far from being accessible to consciousness. In contrast, knowledge that states its predication and factuality explicitly cannot be contained in the use of a procedure. It thus loses efficiency but becomes more flexible, to be used in hypothetical reasoning, evaluation of truth, and conscious awareness. The distinction between procedural and declarative knowledge is a good basis for understanding why voluntary control of action is tied to explicitness and consciousness.

3.4 Voluntary Control.

The dominant philosophical view of what differentiates our intended actions, for which we are responsible, from other movements is that those actions must be caused by our desires and beliefs (Davidson, 1963). Heyes & Dickinson (1993), discussing whether animals act or just respond, argued that intentional action--unlike responses--must be based on an understanding of why one performs them, one must represent the goal one pursues and the fact that the action leads to that goal. Searle (1983) even argued that intentional action must be causally self referential, one must intend that the action be caused by one's intention.

A useful model for this phenomenal distinction between automatic (responses) and controlled, or willed action is that of Norman and Shallice (1980). It distinguishes two levels of control. *Horizontal strands* operate at the level of implementing schemas which consist of complex conditional action tendencies (productions like in Anderson's, 1976, ACT model) with automatic control through activation by triggering stimuli and mutual inhibition of simultaneously triggered schemas (*contention scheduling*). *Vertical strands* of control come from the *supervisory attentional system* (SAS, a close relative of the central executive, Baddeley, 1986). The two control systems are supposed to capture the phenomenal distinction between automatic responses and intentional action as well as explaining why a particular set of actions becomes difficult for patients with problems of voluntary control (e.g., patients with frontal lobe insult). These "SAS tasks" are typically (1) the setting up of new action schemas upon task instructions, (2) monitoring of novel or dangerous actions, or (3) the inhibition or monitoring of interfering action schemas.

Action schemas or productions are complex versions of responses to stimuli. They incorporate procedural knowledge about event contingencies in the world that (as discussed in 3.3) leave predication of these regularities and factuality implicit in their application. The stimuli

15 Another source of inferential limitations that makes for modularityis implicitness of property structure. If there is an inference from 'male' to 'shaves in the morning', it cannot be used on bachelors unless their being male is represented explicitly. So if one domain does not use the same property-structure as, another, even though their concepts overlap the two domains are modular with respect to one another.

that trigger them can be declarative or nondeclarative representations of features of the environment or internal states. The control exerted at the level of contention scheduling as well as that exerted by the SAS is in terms of boosting or inhibiting the activation of schemas. For example, in order to ensure that a single schema produces coherent action the dominant schema might get its activation boosted even further at the cost of the activation of less dominant schemas.

We suggest that contention scheduling directs this control purely on the basis of the schemas as representational vehicles (the amount of activation is a feature of the schema as vehicle, not of its representational content). In contrast, the SAS directs its control on the basis of the schemas' representational content. In support of this contention one can show that such content oriented control is necessary for the 'SAS tasks' listed by Norman and Shallice. In a version of the Wisconsin Card Sorting test for children a three year old child (like a frontal lobe patient) who has learned to sort cards by colour, must now sort the same cards according to a new rule (e.g., the shape of symbols on the card). Without SAS, the oncelearned colour sorting rule is dominant and will suppress execution of the new rule. Three year old children, even though they know the new rule and can verbally state it, will perseverate, sorting according to the old rule (Zelazo, et al., 1995), as frontal lobe patients tend to do on the traditional test (Shallice, 1988). For the SAS to be of use here, it must boost the new schema and inhibit the old, dominant now. This cannot be done on the basis of vehicle features such as amount of existing activation or strength (too many weak schemas would be boosted), the SAS must be able to address the new schema by its content, that stimulus-response sequence that the new rule requires (see Perner, 1998, for discussion of other SAS tasks).

Controling schemas via their content requires the representation of that content. In order to avoid confusion, this content must be explicitly marked as being not factual (i.e., explicit representation of factuality), but something that is desired or intended (explicit representation of attitude). This means that the SAS must be (or contain) a second-order mental state (one that represents desires) which is an important prerequisite (or even a sufficient condition) for being a conscious state according to the higher-order thought theory of consciousness. This analysis hence suggests, that the need to represent content and attitude explicitly distinguishes controlled or willed action from automatic action. We can identify intentional action with action (be it automatic or willed) that is in line with the explicit representations of the SAS. If automatic action contravenes those representations then it is experienced as an unintentional lapse or "slip of action" (Reason & Mycielska, 1982). This analysis also makes it clear why willed action is conscious--because it is based on a second order mental state. With this we have a theoretical justification of why voluntary control is used as a criterion for consciousness in the quite different areas of research on implicit memory and subliminal perception. Note however, that not all aspects of the content of a schema need be explicitly represented to allow control by the SAS; only enough aspects to indicate that the action of the schema is desired. Only those aspects of the content which are explicitly represented will be conscious; the rest may in principle embody knowledge which the person is not aware of having, and whose details of application they could not control. Our argument requires a conscious representation to be made by the SAS (e.g. ‘I want [it to be the case that] that I play Für Elise on the piano’), but the overlap in content between this representation and a body of knowledge (e.g. about piano playing) could allow that knowledge to apply, even if the factuality of the knowledge is not explicitly represented; that is, a fully explicit

representation in the SAS can co-exist with implicit representations in a knowledge base. We will see an example of this in section 4.4 below.

Jacoby's (1991) process dissociation procedure uses voluntary control of knowledge in order to provide better estimates of implicit (unconscious) or explicit (conscious) memory. The procedure can be used not only for memory but also for subliminally presented information (Debner & Jacoby, 1994). One critical part of this procedure is the exclusion condition, in which participants in an indirect test of memory (e.g., to complete word stems) are instructed not to use words that were presented in a list. Unconscious knowledge, in particular, knowledge that leaves predication implicit (e.g., the word form "butter" of the word that was on the learning list), can influence the indirect test and escapes exclusion in the exclusion test, because the word form does not fall under the description "word on that list". So, the number of words from this list that are used as an answer, despite instructions is a better indicator of implicit memory than performance on the indirect test without exclusion instruction, because on the indirect test there is no control for participants using words that they can remember explicitly.[16]

3.5 Summary.

Our analysis of the aspects of knowledge that are represented explicitly and those that are left implicit provides a basis for relating different criteria that have been brought to bear on the implicit-explicit distinction. Knowledge that represents its content, its attitude, and its holder (self) explicitly is on the higher-order thought theory conscious. Explicit representation of factuality might be sufficient, because from being a fact knowledge can be inferred. Explicit representation of predication (and often of factuality) is required to refer in verbal communication and thus a link emerges between direct tests (where reference is made to the known fact) and explicitness and consciousness. Similarly, procedural knowledge leaves predication implicit in its application. It remains accordingly unconscious. Declarative knowledge represents predication and factuality explicitly, thus qualifying for conscious access. Automatic action is based on schemas (productions) that, like procedural knowledge, leave predication implicit, while controlled action (SAS) represents the content of these schemas explicitly together, with the attitude. Willed action is therefore conscious whereas automatic action can remain unconscious. This justifies the use of voluntary control to help distinguish conscious from unconscious elements in task performance.

16Although Jacoby's method constitutes a clear methodological improvement, one must point out a remaining weakness. There is no guarantee that all participants will use the same criterion for excluding information. Consider: Is knowledge that makes predication explicit but leaves factuality implicit (e.g., 'the word "butter" being on the list") sufficient for exclusion? Probably not; it needs to be represented as a fact. But is even that sufficient? Consider the possibility that the origin of this piece of knowledge is not explicitly represented and that consequently no justification for one's judgement can be given; then people under justification pressure, unsure of their intellectual competence, might not consider it a reliable fact and not bring it under the exclusion criterion. In sum, although Jacoby's procedure undoubtedly provides a methodological advance in dissociating implicit from explicit memory, it still suffers from the ambiguities inherent in indirect and direct tests as measures of implicit and explicit knowledge. We will briefly return to the issue of resolving such ambiguity in our discussion of intentional control of knowledge of artificial grammars.

4. Outline of Potential Application to Research Areas.

4.1. Visual Perception.

Visual information is not processed in a unitary way. At least two functionally different systems exist. Traditionally it was thought that the functions were for the perception of objects and the perception of the spatial relations between these objects ('What' versus 'where', Ungerleider & Mishkin, 1982). Recently, Milner & Goodale (1995) have moved from a distinction in terms of encoding different aspects of the visual array to either forming a perceptual representation ('what' there is) versus exerting visuo-motor control ('how' to act). This reconceptualisation has been prompted in large part by functional dissociations in brain injured patients and normal people (e.g., Milner & Goodale, 1995; Rossetti, 1998). As one example we describe a series of experiments by Bruce Bridgeman on the induced Roelofs effect.

Bridgeman (1991, Bridgeman, Peery & Anand, 1997) reports that for human observers a stationary dot within a rectangular frame appears to move opposite to a movement of the frame. After a brief exposure to this apparent movement, the display vanished and the observer had either to indicate verbally at which of five marked locations the dot had been after the movement or to point to the location of the dot. In their verbal responses all observers were susceptible to the illusion and reported the dot's last location as having moved opposite the frame's movement. In contrast, only half the observers were susceptible to the illusion in their pointings; the other half pointed quite accurately to the dot's actual location. Bridgeman interprets the results as showing the dissociation between a cognitive (perceptual) system used for verbal report and a system for visuo-motor control that steers the pointing finger.

This interpretation can be refined within our conceptual framework. Visually guided behaviour can be procedural and nondeclarative, it doesn't need a distinction between facts and non-facts. It is a system that registers object features in egocentric space and everything which is represented is a fact. An interesting question is whether predication needs to be represented explicitly. It seems that the object one grasps does not need to be represented as a re-identifiable individual. Representation of its visible features suffices[17] as Campbell's (1993) analysis shows that orienting oneself in relation to landmarks can be done in a pure feature placing system without the necessity of conceptualising the landmarks as physical objects that have those features. So, no predication of the visible features to the objects that have them needs to be represented. This still leaves the question, however,of whether the visible object features need to be predicated of the spatial positions, i.e., "dot-ness in position x, y, z" which amounts to predicating the feature 'dot-ness' of that position. Or is it sufficient simply to have a conjunction of feature and position? A plausible answer might be that a mere conjunction is sufficient if only a single object needs to be tracked. Then the predication of feature to position can remain implicit

17 To claim that visually guided action can be based on predicating implicit representation may be too radical as Evans (1975) has shown how limited linguistic communication would be without predication. However, visual perception of and action in one's immediate surroundings may be different because relations with in one's egocentric space are much more constrained than relations between linguistically communicating partners. In Campbell's (1993) words, this is possible because the features can be used in a causally indexical way which linguistic communication cannot exploit to the same degree because people typically do not stand in exactly the same causal relation to what they communicate about.

in the tracking. For keeping the position of a second feature in mind while tracking the first, explicit predication is required. We know of no data that speak to this issue[18] , but the question of whether visually guided action leaves only factuality and time or also predication implicit is testable.

In contrast to visually guided behaviour, to give a verbal response is to make a judgement, that that's where the dot really is. The information in this system needs to represent predication and factuality explicitly. As these are preconditions for consciousness, this explains why the information used for the verbal response is what is consciously experienced. The analysis also makes clear a certain ambiguity in the pointing condition. Pointing is a movement of the finger to the target (a visually guided movement); but it is also a declarative act that states what is the case. The bimodal distribution could be due to this ambiguity. From our analysis it follows that if the instructions are not to point but to move one's finger to touch the dot, then no observer should be susceptible to the Roelofs effect. Bridgeman (personal communication) carried out this condition and obtained the predicted results.

Bridgeman's experiment illustrates the other interesting property of the visuo-motor system: that its information persists for only a few seconds. When the response is delayed for eight seconds, all observers show the Roelofs effect just as in their verbal response (and this also holds for the condition where observers had to move their finger to the target, Bridgeman, personal communication). Representations that do not mark factuality and time are only useful to represent the here and now, as they do not differentiate what is a fact (here and now), what is not a fact but mere a hypothetical assumption, or what was a fact but is no longer (see Perner, 1991, for developmental convergence of the ability to represent hypothetical scenarios and to represent change over time). So, because the visuo-motor system leaves time and factuality implicit, it can only update its information about the current state of the environment but cannot keep track of past states of affairs and compare them with the one. For this, factuality and time must be represented explicitly (see alsoWong and Mack 1981).

In sum, what these results demonstrate is that there are two visual information processing systems. One is identified neurophysiologically with the dorsal path from the primary visual cortex (V1) to the posterior parietal cortex (Milner & Goodale, 1995). Its information is unconscious, it cannot be used for statements (verbal or gestural) about the world, it is not susceptible to certain illusions and it is used for action in the world but is of limited duration. Our interpretation is that this system leaves factuality and time implicit (and perhaps also predication--see above). The other system is identified with the ventral path from V1 to the inferotemporal cortex. Its information is conscious, susceptible to illusions and used for statements about the perceived world and for action in the world after some delay. This system represents predication and factuality explicitly and thus makes its content accessible to consciousness. (see also Aglioti, DeSouza, and Goodale, 1995, Gentilucci, Chieffi, and Daprati , 1995, Milner & Goodale, 1995, chapter 6; Rossetti, 1998).

---

18 There may be relevant evidence from subliminal perception for which unconscious perception of the meaning of single words is possible but the subliminal perception of the meaning of word combinations is difficult to demonstrate (Greenwald, 1992; Kihlstrom, 1996) – perhaps because the interpretation of combinations requires explicit predication.

The spared capacities in blindsight and numb-sense patients (tactile analogue to blindsight, Paillard, et al., 1983) depend on similar parametric variations. Marcel (1993) reported that blindsight patient G.Y. was able to detect an illumination change in the blind field better when the response was made quickly than when it was delayed by 2 or 8 seconds, when the response consisted of an eye blink (interpretable as a nondeclarative response) than a verbal "yes-no" (a declarative comment), and when the patient was invited to guess than when instructed to give a firm judgement (where bona fide responses require judgement explicit representation). Marcel also found that people of normal vision responded to near-threshold changes in illumination in the same way as blindsight patients. That is, in people with normal vision, detection was better when responses consisted of an eye blink rather than a ”yes-no” verbal response, and when people were invited to guess rather than make a firm judgement.

A particularly interesting point about the last result is that the response shift from judgement to the guessing condition consisted not of a criterion shift to saying "seen" more often, but of an increase in discrimination accuracy (an increase in hit rate and decrease in false alarm rate). A shift in criterion towards "seen" responses would be expected if the stimulus was encoded explicitly as a fact about which one is uncertain in one's judgement. Then being given leave to guess would simply lower the rejection criterion resulting in an increase in the willingness to say "yes". In contrast, when a stimulus is encoding a fact implicitly, there is a representation "illumination change" but no information as to whether or not it occurred, or whether it occurred on the current or an earlier trial. Thus, there is no proper information for a judgement (hence low detection accuracy). With leave to guess, however, one is free to let oneself be influenced by the fact-implicit information that happens to be correct, which results in higher detection accuracy.

4.2. Memory.

Memory has many different facets. To help focus our discussion we distinguish the wider use of memory as the availability of information acquired in the past (e.g., remembering/ still knowing that 2×2=4) from the narrower meaning of memory as the availability of information about events in the past acquired in the past. As a concrete example, we use the typical memory experiment in which one is read a list of words, among them the word "butter" and we look at the consequences when various aspects of this event are represented explicitly or left implicit. The consequences we consider are in terms of memorial state of awareness, retrieval volition, and test responses.

As the first possibility, we consider strong implicitness. At learning, the word "butter" designed to represent the fact that ”the word ‘butter’ occurred on the list” is stored so that only the word form "butter" is represented explicitly and all the rest is left implicit. This supports no particular memorial state of awareness. It could support a 'feeling of familiarity', if that word had been encountered the first time on that list. This representation cannot be accessed voluntarily, and is not used *bona fide* in any direct test, since no reference to any particular occurrence can be made. It can influence indirect tests, however. The mere presence of the word form "butter" can enhance the likelihood of answering a request to list dairy products with "butter". It could also account for participants reporting ‘butter’ on an exclusion test without any accompanying feeling of familiarity (Richardson-Klavehn, Gardiner, & Java, 1994).

It is also likely that there are cases where it is not just the word form ”butter” that has been represented, but also the perceptual details by which that word form was perceived. That is, a representation of the conjunction of various contextual features is formed, but this feature-complex need not be predicated as having occurred on the list. Such a representation could enhance perceptual identification and produce familiarity effects without supporting recollection (e.g. Jacoby & Dallas, 1981). Such a representation could also be involved in the ”mere exposure effect”, in which exposure to a stimulus, for example a novel shape, can lead to high affect ratings for the stimulus in the absence of recollection of having seen it before (Zajonc, 1968; Bornstein, 1989; Gewei and van-Raaij, 1997).

When the occurrence of the word "butter" is explicitly predicated, i.e., "the word 'butter' occurring on that list," then it can come under direct voluntary control as now reference to the particular event of being on the list is possible. As a consequence, performance on a direct test can be better than on an indirect test (Reingold and Merikle’s, 1993, control for differences in test sensitivity). However, voluntary control remains an educated guess and does not result from a considered judgement, because the occurrence is not represented as a fact.

Explicit representation of the occurrence as a fact makes the event accessible under the description of being a fact and participants can now give a considered judgement that the word ”butter” is part of that list. With explicit representation of time, participants can then also judge that ”butter” occurred at a particular reading of the list in the past. They can experience memory of a past event. It can be a conscious experience of memory of the past according to the higher-order-thought theory, because explicit representation of factuality entails a higher order thought about one’s knowledge. However, even with such a representation participants may remember no details of seeing/hearing the item.

An important next step comes with explicit representation of the experiential source of one’s knowledge: ‘I know that "butter" was on the list because I saw it there’. Only such encoding—encoding of having been in direct contact with the known event—constitutes *genuine episodic memory* according to Tulving (1985; Perner, 1991).[19] Tulving (1985; and later others, such as Gardiner, 1988) distinguished two types of recognition responses: those accompanied by simply an experience of **K**nowing that the item occurred earlier in the context of the experiment (”K” responses); and those based on truly **R**emembering the prior experience of the item (”R” responses).

”K” responses may arise for various reasons, e.g., because the word form ‘butter’ is encoded predication implicitly and simply comes to mind readily (whether the participant does give a positive recognition response depends on his theory of why the word came to mind) or because a predication explicit representation has been formed and hence the participant guesses that the word had been on the list. In both cases, the participant may give a ”K” response with low confidence. On the other hand, if the participant experiences strong familiarity when he comes across the word ”butter”, he may give a ”K” response with strong confidence. However, in all

19 Dokic (1997) pointed out that the above formulation of the memory trace still leaves room for counterexamples. In order to ensure a true episodic memory the encoding has to be self-referential in Searle’s (1983) sense: ‘I know that (”butter” was on the list and this knowledge comes directly from my past experience of the list)’. The parenthesis are added to bring out more sharply the syntactic embedding that makes ”this knowledge” self-referential.

these cases there is no genuine knowing that ”butter” was on the list, just guesses that carry more or less conviction. Researchers in the field (Conway et al, 1997) have now started to give participants a choice between ”K” responses and ”guesses”. This may separate predication and fact implicit knowledge from knowledge that represents the factuality (and past-ness) of the event in question explicitly. Unlike ”guesses,” "K” responses should not just be produced but produced as the reflection of a fact.”R” responses differ from ”K” responses in that they need be seen not only as reflecting facts but also as products of one’s direct experience.

Table 2 summarises the different levels of explicitness, and the memorial state of awareness, voluntary control and kind of test performance they support. Our analysis yields distinctions that map reassuringly onto distinctions that have emerged from the empirical literature. In particular, it can address the distinction between *retrieval volition* and *memorial state of awareness* (Richardson-Klavehn, Gardiner & Java, 1996; Schacter, Bowers, & Booker, 1989); it honours the distinction between ”implicit” memory and the distinction between "know" and "remember" judgements as two kinds of explicit memory in the spirit of Tulving’s (1985) original distinction, where "know" judgements are supposed to cover 'knowledge of the past' and "remember" judgements memories of experienced events as experienced (Perner, 1990). This analysis indicates that both ”R” and ”K” count as declarative knowledge (both involve explicit predication) and familiarity can be purely procedural (predication left implicit).

Insert Table 2 about here

4.3. Development.

In our framework there is no simple dichotomy between implicit and explicit knowledge. This owes much to Karmiloff-Smith's (1986, 1992) insistence that the basic dichotomy should be embellished by further levels of explicitness. It is reassuring that our framework unfolds logically from the conceptual analysis of knowledge and yields a plausible correspondence to Karmiloff Smith's empirically motivated classification. Her initial level (I) of implicit knowledge, where the information is only <u>in</u> the system maps onto procedural knowledge, which leaves predication implicit. Her first level (E1) of explicit knowledge results from a redescription of the original information encoded in procedural format, so that the information becomes information <u>to</u> the system, useable by different parts of the system. This maps onto knowledge that makes predication explicit (and can thus be referenced flexibly by different user systems) but leaves factuality implicit. At the next level of ”explicitation” (E2) the knowledge becomes conscious, and at the final level (E3) it also verbally expressible. The once clear progression from E2 to E3, has later been collapsed into a level E2/3 (1992, p. 23) owing to the lack of a clear empirical demonstration of such a progression. The level E2/3 corresponds to knowledge that makes factuality (and source) explicit. Moreover, since explicit factuality tends to make knowledge conscious and verbally accessible, our analysis actually suggests the merging of the original levels 2 and 3.

Whereas Karmiloff-Smith's research emphasises how implicit knowledge becomes

increasingly explicit with development, dissociations between two competing knowledge bases have also been found—dissociations reminiscent of those in visual perception (e.g. Diamond & Goldman-Rakic, 1989; Goldin-Meadow, Alibali, and Church, 1993; Clements & Perner, 1994). Goldin-Meadow et al review studies that show that the acquisition of concepts of quantity (Piaget & Inhelder, 1974/ 41) can be more advanced in children's gestural comments than in their verbal responses. One of the interpretations of this finding was (Church & Goldin-Meadow, 1986) that the multidimensional spatial medium of hand gesture makes it easier to express novel ideas than the unidimensional temporal medium of linguistic expression. However, one can think of the gestures as spontaneous (mostly unconscious) concomitants of the thinking process. In that case the earlier emergence of advanced knowledge might be the sign of thoughts about reality that have not yet been recognised as being about reality (implicit factuality). This interpretation fits a parallel finding in children's developing "theory of mind".

Clements and Perner (1994) reported that the understanding of false belief emerges in children's visual orienting responses as early as 2 years and 11 months, a year earlier than in their verbal responses to questions. Children are told enacted stories in which the protagonist does not see how his desired object is unexpectedly transferred from one location(A) to another (B). Children in the interesting period around 3 years of age answer the question about where the protagonist will go to get his object wrongly by pointing to the current location of the object. However, a majority of these children look (visual orienting responses) in anticipation of the protagonist at the empty location where the protagonist mistakenly thinks the object is.

Further research (Clements & Perner, 1996) indicates a remarkable similarity to the observed dissociations between the two visual systems (see Section 4.1). When instructed to move a welcoming mat for the mistaken story protagonist who was on his way to get his object, children who move the mat spontaneously tend to move it correctly to where they think the object is (A), whereas children who need prompting (thus with some delay) move it to where the object is (B). There seems to be a stage in children's developing understanding of belief where two different knowledge bases dissociate. One of them is a more accurate and developmentally advanced knowledge base (by analogy with the dorsal visual path) that supports only nondeclarative action (looking and moving a mat) carried out without delay (spontaneous mat move) while a less accurate and less developmentally advanced knowledge base (analogous to the ventral visual path) is used for declarative responses (verbal and pointing) and delayed action (prompted mat moving). We do not know, of course, whether the more advanced knowledge is conscious and the other unconscious, since one cannot ask 3 year old children to report on such a distinction, but otherwise the similarities are remarkable.

Such a similarity between dissociations in processing visual information about the environment and understanding another person's false belief suggests that the characteristics of the two types of knowledge are not determined primarily by the brain regions in which the information is processed (dorsal vs. ventral path) but by more general functional differences that apply to visual information processing as well as a theory of mind. Our analysis shows how these functional distinctions could arise from these aspects of knowledge thatare represented explicitly. An interesting speculation about functional differences in the theory of mind case is that the explicit understanding comes with (something of) a real theory, i.e., a causal understanding of belief formation and how belief determines action. In contrast, the implicit understanding of

where the protagonist will go may be based on abstraction of situational regularities. Within our framework this assumption gives a quite coherent picture of the existing data and leading to new, testable predictions (Perner & Clements, in press).

One can learn that certain events tend to go together and form a typical sequence. Such filtering of statistical patterns of possible combinations does not need representation of individual events and inferences from individual events to all possible events. Rather it is a process of pattern formation and recognition for which connectionist systems are good (e.g., to classify different feature patterns into letters, e.g., Bechtel & Abrahamsen, 1991). The combinations of letters encountered in artificial grammar tasks have a similar effect and can be particularly well modelled by connectionist networks (Dienes, 1992). Although individual instances shape the connections between units and hence the association between the properties these units represent, there is no representation of the individual instances.[20] Connectionist work also shows that such pattern generalisation leads to pattern completion. If many elements of a typical pattern are present then the network tends to generate representations of the missing bits. This is important, because such pattern completion processes can produce expectations of what is to come on the basis of what has happened so far. For us, the important implication is that such associative expectation is possible without explicit predication.

This makes it possible to anticipate correctly where the protagonist will go to get the desired object in our false belief stories without explicit predication to a particular occasion, i.e., without representing *that* he will go there. According to our discussion, such a representation of the mere event form 'protagonist going to location A' and hence, 'protagonist at location A' as part of a pattern completion process can guide visual orienting responses and spontaneous actions because it can trigger an existing action schema waiting to be executed. It cannot be used for communication because it fails to be predicated of an individual event which can be re-identified across mental spaces explicitly marked as, "facts", "anticipation", or "verbal description." It cannot sustain uncertainty, as it does not support a self-reassuring check about where the protagonist *will* come down because without explicit predication there is no representation stating *that* he will go anywhere. This is the pattern of results we observed in the precociously correct responses: they were high only in spontaneous action and visual orienting responses.

In contrast, a theory of belief goes beyond mere generalisations of observed regularities and constitutes genuine causal understanding of the underlying processes (see Gopnik, 1993; Perner, 1991, for indications of theory use). Causal understanding cannot be achieved by mere pattern matching and pattern completion but must use explicit predication because causal reasoning is supports counterfactuals (Lewis, 1986; Salmon, 1984). Counterfactual support means one understands that if the conditions had been different, the result would have been different; such reasoning requires different mental spaces for contrasting the actual facts with their counterfactual oppositions. For these reasons, responses based on a causal theory of belief should

20 For this reason one can speak of association but not of inference.Inferences go from state of affairs to state of affairs, that is, reasoning of the form 'whenever X is the case then Y must be the case.' But that means X and Y are predicated of particular occasions. That associative processes but not inferences are possible implicitly and without consciousness is reminiscent of Sloman's (1996) suggestion that implicit knowledge is tied to associative processes and explicit knowledge to rule governed inference processes.

also be accessible to communication (answers to questions) and be robust against doubt (hesitating action).

One, accordingly, can predict that implicit knowledge should be shown primarily in the situation described above, where the correct response can be based on situational, behavioural regularities, such as "people look for objects where they last put them, where they last saw them, where they told someone to put it, etc.". In the traditional scenario all these regularities—if they apply—point to the same, correct answer "A". In a variant scenario (Perner, Leekam & Wimmer, 1987) the protagonist, who has put the object into B, tells a friend to move the object from B to A, but the friend forgets. Here, behavioural regularities give different predictions. "Last seen" or "where put" indicate location B while "told to put" indicates A correctly. Hence signs of implicit understanding should be reduced in this scenario. Indeed, Clements (1995, Chapter 5) reports that children show fewer orienting responses to location A than in the traditional scenario. In contrast, their verbal responses show little difference in the two scenarios, replicating the original result by Perner et al. (1987). This is to be expected if explicit responding is based on a causal understanding of belief formation.

Another prediction is that verbal explanations of why the protagonist believes the object is still in location A (in the original scenario) in contrast to observing behavioural regularities (seeing the protagonist look for the object in A) should affect implicit and explicit understanding differently. Causal explanations should primarily affect explicit understanding ,whereas observing regularities should have a stronger effect on implicit understanding. The role of explicit understanding of this prediction has been tested. Clements, Rustin & McCallum (1997) report that causal explanations affect verbal responses but the observation of regularities does not. The corresponding data on visual orienting responses or action responses are not yet available.

4.4 Artificial grammar learning

Our framework also elucidates the different ways in which knowledge can be implicit in the standard implicit learning paradigms. The paradigm explored most thoroughly in the implicit learning literature is artificial grammar learning (see Reber, 1989, and Berry, 1997, for overviews). In a typical study, participants first memorize grammatical strings of letters generated by a finite-state grammar. Then they are informed of the existence of the complex set of rules that constrains letter order (but not what they are), and are asked to classify grammatical and nongrammatical strings. In an initial study, Reber (1967) found that the more strings participants had attempted to memorize, the easier it was to memorize novel grammatical strings, indicating that they had learned to use the structure of the grammar. Participants could also classify novel strings significantly above chance (69%, where chance was 50%). This basic finding has now been replicated many times. So participants clearly acquire some knowledge of the grammar under these incidental learning conditions, but is this knowledge implicit? We will now analyze the case of artificial grammar learning theoretically and empirically in terms of the different aspects of being a fact or being knowledge that can be made explicit, or left implicit, according to our previous analyses. (See also Dienes and Perner, 1996, who explore whether participants represent the property structure of a grammar implicitly or explictly, an issue not dealt with in the following.)

4.4.1 Predication

When participants learn the structure of an artificial grammar by exposure to the exemplars, they may not explicitly represent the particular grammar to which the properties are predicated. Consider a person who uses the mental rule that "M can be followed by T". This statement represents the fact that, according to the grammar one was trained on 10 minutes ago, M can be followed by a T. Yet, the fact that it is a particular grammar which has this property is not explicitly represented because there is nothing in the expression "M can be followed by T" whose function it is to covary with that fact. This fact can be made explicit by forming the mental expression: "g has the property that M can be followed by a T", where g denotes a particular grammar (e.g., the grammar that I was just being trained on). The critical feature here is that different properties, such as "my having just been trained on" and "being a grammar in which M can be followed by T" can both be predicated of g. This extended expression makes the implicit predication of 'M is followed by a T' of a particular grammar explicit, because the whole expression does have the function of covarying with the fact that the identified particular grammar is characterized by the property in question.

Whether participants represent the individual grammars and the predication relationship explicitly can be revealed by the *volitional control* that participants have over the application of their knowledge. Consider a test of volitional control given to participants by Dienes, Altmann, Kwan, and Goode (1995). Participants were given 7 minutes to try to memorize exemplars generated by one grammar, and then another 7 minutes to try to memorize exemplars involving the same 6 letters generated by a second grammar. Participants were then informed that two grammars were involved and given a test in which a third of the items followed the first grammar (but not the second, e.g.: xmxrtvtm), a third followed the second grammar (but not the first, e.g.: xmvrxrm), and a third violated both grammars (e.g.: xmtvvxrm). Participants were asked to choose items that followed only one of the grammars; half the participants were asked to endorse only the items consistent with the first grammar, the other half only the items consistent with the second grammar. Participants were perfectly able to distinguish the grammars at the usual performance level in such tasks and showed no tendency to endorse the grammar they were asked to ignore. How could this performance be achieved?

One way to succeed in such a test is to have direct volitional control over one's knowledge, in the sense that one can decide to use or not to use it because it has been explicitly labelled as the particular body of knowledge one wishes to use or not use. That is, we assume that for direct control it is necessary to represent the individual grammar explicitly. There are other ways of controlling which body of knowledge to use, however, that do not require such explicitness. For example, Whittlesea and Dorken (1993) argued that participants could distinguish different grammars by familiarity. One account of the Dienes et al (1995) results along these lines is that the choice of grammar can be made by means of a compound property (e.g., in-context-A,-M-can-follow-T). Context A could be, for example, a particular time at which a string was studied . If context A is reinstated by task demands or imagination, the knowledge of a particular grammar can be isolated (through association) without having to predicate these properties explicitly to any particular grammar.

Even though this scenario of indirect control over particular grammars without explicit representation of the grammar is often possible or even plausible, there may be situations in which one can plausibly decide that volitional control was actually mediated (at least in part) by explicitly representing the individual grammar. For example, if, with a sufficiently sensitive test, measures of familiarity (such as ratings, speed of stimulus identification) do not predict classification response, then these alternative scenarios (that do not represent the individual explicitly) are not supported. Buchner (1994) in fact found that grammaticality judgements were not related to speed of identification. If this type of observation is supported, it follows from the volitional control experiments that participants do represent the individual grammar (and the predication relationship) explicitly. Of course, as we noted earlier (Section 2.1.3), the presence of knowledge in which the predication relationship is represented explicitly does not rule out the possibility that there is further knowledge on the same topic which is predication implicit.

4.4.2 Reflection on Attitude

To clarify how explicitly participants can reflect on their knowledge it is necessary to be clear about *what* piece of knowledge participants may be reflecting on (e.g., Shanks & StJohn's, 1994, information criterion). We distinguish two different domains of knowledge. The first we call grammar rules. These are the general rules of the grammar that the participant has induced; e.g. "M can be followed by T". The second domain pertains to the ability to make grammaticality judgements. This arises when the grammar rules are being applied to a particular string and it pertains to the knowledge of whether one can judge the grammaticality of the given test string independently of any knowing that one knows the rules one brings to bear for making this judgement.

Knowledge of artificial grammars and of natural language may differ. We seem to lack explicit knowledge of the grammar rules both of English (we cannot represent *any* sort of attitude towards most rules of English grammar, so such rules are at least attitude implicit) and of the quickly acquired artificial grammars. In contrast, we are fully aware and have explicit knowledge of our ability to judge the grammaticality of English sentences. We lack this sort of explicit knowledge of our ability to judge the nonsense strings produced by an artificial grammar. We may also lack it in the early stages of learning a first or second language as well).

Various relationships between the knowledge of rules and grammaticality judgements are possible. Reber (e.g.1989) showed that people do not use the rules to respond deterministically; that is, when retested with the same string, participants often respond with a different answer. Extending this argument, Dienes, Kurz, Bernhaupt, and Perner (1997) argued the data best support the claim that participants match the probability of endorsing a string as grammatical to the extent to which the input string satisfies the learned grammatical constraints, and that this probability varies continuously between different strings. Learning increases the probability of saying ”grammatical” to grammatical strings and decreases it for nongrammatical strings. As people begin to learn, the probabilities start to covary with the success, with a higher probability of correctly identifying strings that actually are grammatical. This means that the probabilities actually imply the epistemic status of the grammaticality judgement, ranging from a pure guess to reliable knowledge. The probabilities have the function to capture this information, because without this correlation the system would not be successful and the relevant learning mechanism

would not have evolved. However, the mechanism responsible for producing these probabilities need not explicitly represent that there is knowledge (i.e. that the representations induced by training and testing have the properties given in section 2.1.2). For example, there is no need for the mechanism to represent that there is something that is taken as reflecting the accuracy of the judgements, nor that the accuracy of the judgements is well-founded in the learning history, nor that the self is the possessor of the knowledge.

Although participants' response probabilities suggest only a structure-implicit representation of the accuracy of their judgements, we do not know whether they have a more explicit representation of it. One way to test whether they can represent the epistemic status of their judgements explicitly is to ask them to state their confidence in each classification decision, (e.g. on a scale which ranging from 'guess', through degrees of being 'somewhat confident', to 'know'). If the confidence rating increases with the probability of responding correctly to each item, with random responding given a confidence of 'guess', and deterministic responding given a confidence of 'know', then the propositional attitudes implied by the probabilities have been used by the participant to explicitly represent the epistemic status of the grammaticality judgements; if confidence ratings are not so related to response probabilities, then epistemic status has been represented only implicitly.

The above only tests whether participants represent their ability to make judgements as knowledge. It is possible, as in the natural language case, that they know when they have the knowledge for judging grammaticality and when they are guessing, but still their knowledge of grammar rules is not represented as knowledge. This could be tested if we knew the actual content of participants' grammar rules. If the rules have been induced over time by some kind of optimal learning rule, then the epistemic status of the rules must be greater than just guesses. If participants, despite stating rules freely, or endorsing presented rules, nevertheless believe they are just guessing, then the rules have not been appropriately represented as knowledge. Also, if the rules had not been represented as knowledge, they may not be offered as descriptions of the grammar, because participants would not know that they knew anything. Of course, failure to state the rules in free report could also arise for other methodological reasons owing to the normal failings of free recall.

Establishing whether participants represent knowing their grammaticality judgments or grammar rules explicitly or implicitly is methodologically easier; the relevant research to date has focused on judgements. As noted above, one way to determine whether participants explicitly represent their ability to make judgements as knowledge would be to determine for each test item the probability with which it is given the correct response. If a plot of confidence against probability is a monotonically increasing line going through guess (0.5) to know (1.0) then participants have fully used the implications of the source of their response probabilities to infer an explicit representation of their state of knowledge. If the line is horizontal, then their knowledge is represented purely implicitly. If the line has some slope, but participants perform above chance when they believe they are guessing, then some of the knowledge is explicit and some of the knowledge is implicit.

In artificial grammar learning experiments, participants typically make one or two responses to each test item so it is not possible to plot the confidence-probability graph just

described, but it is not strictly necessary to do so. Consider the case where the participant makes just one response to each test item. We divide the items into those with which the participant makes a correct decision ('correct items') and those with which the participant makes an incorrect decision ('incorrect items'). If accuracy is correlated with confidence, the correct items should be a selective sample of those given a higher average confidence rating than the incorrect items. Conversely, if participants do not assign greater confidence to correct than incorrect items, then that is evidence that the slope of the graph is zero; i.e. they do not represent their state of knowledge of their ability to judge correctly. If participants give a greater confidence rating to correct than incorrect items, that is evidence of at least some explicitness. If in this case, participants perform above chance when they believe they are literally guessing, that is evidence of some implicitness in addition to the explicitness.

Note that the previous paragraph presumes (1) a certain theory of how participants apply their knowledge (probabilistically, rather than deterministically) and (2) that the knowledge is largely valid. Reber (1989) has consistently argued that people's incidentally acquired knowledge of artificial grammars is almost entirely veridical. If people had applied partially valid rules deterministically, there would be no difference between confidence in correct and incorrect decisions, irrespective of whether the knowledge was attitude explicit. Thus, applicying the procedure in different domains requires careful considering how knowledge is applied in each.

Chan (1992) was the first to test whether participants explicitly represented knowing their grammaticality judgements. Chan initially asked one group of participants (the incidentally trained participants) to memorize a set of grammatical examples. In a subsequent test phase, participants gave a confidence rating for their accuracy after each classification decision. They were just as confident in their incorrect decisions as they were in their correct decisions, providing evidence that knowing was represented only implicitly. He asked another group of participants (the intentionally trained participants) to search for rules in the training phase. For these participants, confidence was strongly related to accuracy in the test phase, indicating that intentionally rather than incidentally trained participants represented their knowing more explicitly. Manza and Reber (1997), using stimuli different from Chan's, found that confidence was reliably higher for correct than incorrect decisions for incidentally trained participants. On the other hand, Dienes et al. (1995) replicated the lack of correlation between confidence and accuracy, but only under some conditions: the correlation was low particularly when strings were longer than three letters and presented individually. Finally, Dienes and Altmann (1997) found that when participants transferred their knowledge to a different domain, their confidence was not related to their accuracy.

In summary, there are conditions under which participants represent knowing grammaticality implicitly on most judgements, but there is sometimes evidence of having an explicit attitude of knowing. Even in the latter case, there is usually evidence of implicit knowledge: Both Dienes et al. (1995) and Dienes and Altmann (1997) found that even when participants believed they were literally guessing, they were still classifying substantially above chance.

Dienes et al (1995) provided evidence that this type of implicit knowledge was qualitatively different from knowledge about which the participants had some confidence. When

they performed a secondary task (random number generation) during the test phase, the knowledge associated with 'guess' responses was unimpaired, but the knowledge associated with confident responses was impaired (to a level below that of the knowledge associated with 'guess' responses). That is, this criterion is not just another curious way of categorizing knowledge: It may separate knowledge in a way that corresponds to a real divide in nature.

4.4.3 Summary

In summary, when participants learn artificial grammars, there is evidence that for at least some of the acquired knowledge, participants represent the grammar of which the knowledge is predicated and can thus exert intentional control over which body of knowledge to apply. This intentional control indicates, by our analysis in section 3.4, that the participants have conscious knowledge of some content predicated of that grammar - in particular, the content they use to choose the grammar. There is no need to suppose, however, that participants were conscious of any further aspect of their knowledge (e.g., what the rules of their induced grammar were). If, based on task instructions, participants form the representation 'I am thinking that I should apply the first grammar I studied' they are conscious of their desire to apply the first grammar. If the knowledge pertaining to this grammar is represented predication-explicitly, the mental specification that that is the grammar they want to apply may be sufficient to ensure that it does apply, so the participant has volitional control because of the predication explicitness of the representations formed during learning. The representations of the knowledge about the grammar may not make explicit that the rules are facts, however, or that the knowledge is knowledge. In that case, participants may have volitional control but may regard their responses as guesses, an outcome found by Dienes et al (1995). In several studies, there was evidence that participants did not explicitly represent knowing many of their grammaticality decisions, thus they were not conscious of this knowledge as knowledge. The reason for this is precisely that participants did not have conscious knowledge of their grammar rules and hence could not know that their grammaticality decisions were based on sound knowledge.

These comments illustrate how one can empirically tease apart whether or not the knowledge is predication implicit or attitude implicit.This allows future research to determine which aspects of knowledge are left implicit in the representations formed during different types of learning. Such research could address whether different types of implicitness correspond to qualitatively different learning systems. In addition, future research needs to address other implicit learning paradigms (see Dienes & Berry, 1997, and Stadler & Frensch, 1998, for detailed reviews of implicit learning generally.)

5. Conclusion

In this target article, the natural language meaning of the implicit-explicit distinction was applied to knowledge representations, with knowledge taken as an attitude held towards a proposition. A series of different ways in which knowledge could be implicit or explicit followed directly from the approach. The most important type of implicit knowledge consists of representations that merely reflect the properties of objects or events without predicating them of any particular entity. The clearest cases of explicit knowledge of a fact are representations of one's own attitude of knowing that fact. We argued that knowledge capable of such fully explicit

representation provides the necessary and perhaps sufficient conditions for conscious knowledge. This is consistent with Kihlstrom et al's (1992) suggestion that it is bringing knowledge representations "into contact with" the representation of the self that makes consciousness possible, because that connection defines the self as an experiencing agent in possession of the knowledge. Kihlstrom et al suggested that this connection to the self is lacking in implicit perception; we agree, and add that the lack may be even deeper: the perceptual knowledge may lack not only representation of the self, but even predication to a particular event (e.g. what happened a few seconds ago).

Our analysis also corresponds in places to some recent analyses by Cleeremans (1997) and Dulany (1991, 1996). According to Cleeremans (1997) , "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 (p. 199)". Knowledge can be an 'object of representation' if participants can metarepresent their representation of the knowledge as having various properties; for example, if they can metarepresent it as accurate (or inaccurate), as judged to be true (or false or undecided), or as properly caused (or not). Thus, Cleeremans's criterion corresponds to one aspect of the distinction between attitude implicit and explicit; in particular, to whether the metarepresentation (0) (that "the representation of *Fb is a fact* is possessed by the system"[21]) given in section 2.1.2 is formed. If the content of a piece of knowledge, acquired by a reliable process, can be specified by the participant even as a guess, then it is not implicit according to Cleeremans's criterion. As we argued in the section on artificial grammar learning, behaviour may indicate that a grammatical decision has been taken to be accurate (by consistent responding), but the participant may judge the decision to be a guess. Thus, the attitude of knowing implied by the participants' behaviour has not been explicitly represented. The piece of knowledge 'this string is grammatical' is unconscious as knowledge, but it is conscious as a guess, because the participant can entertain higher order thoughts about it ('I guessed that this string was grammatical'). A deeper form of implicitness occurs when one cannot even entertain a higher order thought about the knowledge; this corresponds to Cleeremans's definition of implicit and to complete attitude implicitness in our terminology.

Cleeremans argues that connectionist networks are particularly suitable for producing implicit knowledge, an analysis that agrees with our own (see Dienes & Perner, 1996). In a connectionist network, the only information available for further transmission through the system is the activation of units (by assumption, for a real connectionist network, not a simulated one). Thus, knowledge embedded in weights is simply not available to be represented as accurate or inaccurate knowledge; hence it naturally satisfies Cleeremans's definition of implicit. On the other hand, Cleeremans argues that in a symbol system representations appear to have at least the potential to be attitude explicit because the system that uses them could always decide whether or not it possesses them. Dulany (1996) makes a stronger claim. Like us, he describes consciousness as involving an agent (I) holding an attitude towards some content; but according to Dulany, propositional content is always conscious.

Our analysis makes a distinction between predication explicitness (which could be a symbolic representation 'Fb') and, among other things, explicit representation of attitude; only the latter representation would produce consciousness of the content Fb. It may be true as a matter

21 We previously called this distinction 'content implicit vs explicit' (Dienes & Perner, 1996).

of empirical fact that any predication explicit representation also allows attitude explicitness; then Dulany's claim would be true. This is a bold empirical hypothesis, but our analysis makes clear that there is no a priori reason for believing it to be true - why should a representation formed, for example, for some local need by a part of our perceptual system *inevitably* allow attitude explicit representations? In section 4.1 we indicated that the predication explicitness of some types of (factuality implicit) perceptual knowledge is an open testable question.

Both Dulany (1991, 1996) and Jacoby (e.g. Jacoby, Lindsay, & Toth, 1992) argued that implicit processes change subjective experience (see also Perruchet & Gallego, 1997). In our analysis, predication implicit knowledge (i.e. maximally implicit knowledge) can change behaviour and we take it for granted that such behavioural change is accompanied by conscious experiences. In a subliminal perception experiment, for example, the activation of the word form 'red' may lead to a 'red' response on a forced choice objective test. This behaviour would be accompanied by the thought 'red pops into mind', or something similar. But the perceptual event would not have been consciously experienced as a perceptual event; that would have required the representation 'I am seeing the word red on the screen' (fully attitude explicit knowledge) to be produced directly *by* the act of seeing the word red on the screen. The predication implicit representation 'red' might trigger inferential thoughts to the effect that 'I must have seen the word red on the screen'. These higher order thoughts enable the participant to be conscious of the possiblity of having seen red, but those inferences do not constitute the conscious perception of red. So, like Jacoby, Dulany, and Perruchet we do suppose that implicit knowledge is often accompanied by conscious experience; one must simply be clear about what it is that the person is conscious of. We do not claim, however, that all implicit knowledge leads to conscious experience. The perceptual system could consider various perceptual hypotheses (e.g. predication implicit features, concepts, or schemata) before settling on one (e.g. Marcel, 1983b), predicating it to an individual. The other hypotheses might never influence conscious experience at all (although they had the potential). Also, a representation may not itself lead to conscious experience, but it might cause other representations downstream of processing that produce conscious experience.

Similarly, an attitude implicit rule may lead one to feel good about a particular part of an English sentence or other grammatical string; this is a conscious experience, but not of the rule. A participant implicitly learning an artficial grammar might induce the rule 'T can follow M', without predicating it of a grammar, representing it as a fact, or representing an appropriate attitude towards it. Nonetheless, the knowledge may make the bigram 'MT' look familiar, inducing a conscious experience that 'MT looks natural'. The participant might infer the further thought: 'in this grammar, perhaps T can follow M'. If this happens, the participant, by observing his own behaviour, has induced a piece of explicit knowledge that co-exists with prior implicit knowledge. Within the participant's knowledge box is the unconscious representation 'T can follow M', not predicated of any particular grammar or represented as a fact. In addition, there is in the knowledge box the conscious representation 'I see that MT looks natural'. Sometimes the unconscious and conscious representations will contradict each other, as in the experiment by Bridgeman (1991) reported in section 4.1.

Our analysis of the meaning of implicit is in itself neutral on the question of whether different systems are responsible for producing knowledge of different degrees of implicitness.

However, different degrees of implicitness will be useful for different purposes, and our view of the evidence is that different systems often do realize different degrees of implicitness in their knowledge (for example see Section 4.1). Dienes and Berry (1997) reviewed the field of implicit learning and concluded that there was a natural divide between learning that produced knowledge about which participants did or did not explicitly represent the attitude of knowing (as we indicated in section 4.4 on artificial grammar learning). Dienes and Berry recommended picking out attitude implicit knowledge using confidence ratings, looking at whether participants performed above chance when they claimed they were just guessing. This "guessing criterion" was found to be useful in separating types of knowledge that were qualitatively different in other respects (e.g. guessing knowledge was found to be resistant to secondary tasks as compared to knowledge about which participants had confidence); but it is still a testable empirical question whether it is attitude implicitness/explicitness that distinguishes different learning systems. We suggest that implicit learning is a type of learning resulting in knowledge which is not labelled as knowledge by the act of learning itself. Implicit learning is associative learning of the sort carried out by first-order connectionist networks (Clark & Karmiloff-Smith, 1993; Shanks, 1995; Dienes & Perner, 1996; Cleeremans, 1997). Explicit learning is carried out by mechanisms that label the knowledge as knowledge by the very act of inducing it; a prototypical type of explicit learning is hypothesis testing. To test and confirm a hypothesis is to realize why it is knowledge. Participants in an implicit learning experiment are quite capable of analyzing their responses and experiences, drawing inferences about what knowledge they must have. These explicit learning mechanisms, when applied to implicit knowledge, can lead to the induction of explicit knowledge. As a result, the guessing criterion is an *imperfect* (but still informative) guide for picking out implicit knowledge; it is not the guessing criterion but the nature of the underlying representations that defines the knowledge as implicit.

In summary, we have presented a framework that makes clear the precise ways in which knowledge can be made implicit. It indicates *why and how* various notions such as consciousness, verbalizability, volition are related to each other and to the notion of explicit knowledge. It also suggests testable predictions about cognitive development, vision, learning, and memory.

References.

Aglioti, S., DeSouza, J. F. X., & Goodale, M. A. (1995). Size-contrast illusions deceive the eye but not the hand. *Current Biology*, 5(6), 679-685.

Anderson, J.R. (1976). *Language, memory and thought.* Hillsdale, NJ: Erlbaum.

Armstrong, D. (1980). *The nature of mind and other essays*. Ithaca: Cornell University Press.

Baddeley, A. (1986). Modularity, mass-action and memory. Special Issue: Human memory. *Quarterly Journal of Experimental Psychology Human Experimental Psychology, 38,* 527-533.

Barwise, J. & Perry, J. (1983). *Situations and attitudes.* Cambridge, MA: MIT Press

Barwise, J. (1987). Unburdening the language of thought. *Mind & Language*, 2, 82-96.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind: An Introduction to parallel processing in networks.* Oxford, England: Basil Blackwell Inc.

Berry, D. C. (Ed.) (1997). *How implicit is implicit learning?*. Oxford: Oxford University Press.

Block, N. (1994). Consciousness. In S. Guttenplan (Ed.), *A companion to the philosophy of mind.* (pp. 210-219). Oxford: Basil Blackwell.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18,* 227-287.

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research 1968-1987. *Psychological Bulletin, 106,* 265-289.

Bridgeman, B. (1991). Complementary cognitive and motor image processing. In G. Obrecht & L. W. Stark (Eds.), *Presbyopia research: From molecular biology to visual adaptation* (189-198). New York: Plenum Press.

Bridgeman, B., Peery, S., & Anand, S. (1997). Interaction of cognitive and sensorimotor maps of visual space. *Perception & Psychophysics, **59**,* 456-469.

Buchner, A. (1994). Indirect effects of synthetic grammar learning in an identification task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 550-566.

Campbell, J. (1993). The role of physical objects in spatial thinking. In N.Eilan, R. McCarthy & B. Brewer (Eds.), Spatial representation (65-96). Oxford: Blackwell.

Carruthers, P. (1992). Consciousness and concepts. *Proceedings of the Aristotelian Society, Supplementary Vol. LXVI,* 42-59.

Carruthers, P. (1996*). Language thought and consciousness. An essay in philosophical psychology*. Cambridge: Cambridge University Press.

Chan, C. (1992). Implicit cognitive processes: theoretical issues and applications in computer systems design. Unpublished D.Phil thesis, University of Oxford.

Cheesman J. & Merikle, P. M. (1984). Priming with and without awareness. *Perception & Psychophysics*, 36(4), 387-395.

Cheesman J. & Merikle, P. M. (1986). Distinguishing conscious from unconscious perceptual processes. *Canadian Journal of Psychology*, 40(4), 343-367.

Church, R.B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an indeof transitional knowledge. *Cognition, 23,* 43-71.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: A psychological and philosophical perspective on the development of thought. Mind and Language, 8, 487-519.

Cleeremans, A. (1997). Principles for implicit learning. In D. Berry (Ed.), *How implicit is implicit learning?* (pp 195-234). Oxford: Oxford University Press.

Clements, W.A. (1995). Implicit theories of mind. Unpublished doctoral dissertation, University of Sussex.;

Clements, W. & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, **9**, 377-397.

Clements, W. A. & Perner, J. (1996). Implicit understanding of belief at three in action. Unpublished manuscript, University of Sussex.

Clements, W.A., Rustin, C., & McCallum, S. (1997). Promoting the transition from implicit to explicit understanding: A training study of false belief. Unpublished manuscript, University of Sussex.

Conway, M.A., Gardiner, J. M., Perfect, T.J., Anderson, S.J. & Cohen, G.M. (1997) Changes in memory awareness during learning: The acquisition of knowledge by Psychology undergraduates. *Journal of Experimental Psychology: General, 126*, 393-413.

Cosmides, L. (1989). The logic of social exhange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, **31**, 187-276.

Currie, G. (1982). *Frege, an introduction to his phylosophy.* Brighton, Sussex: The Harvester Press Limited.

Currie, G. & Ravenscroft, I. (in press). *Meeting of minds: Thought, perception and imagination.* Oxford: Oxford University Press.

Dagenbach, D., Carr, Th. H. & Wilhelmsen, A. (1989). Talk-induced strategies and near-threshold priming: Conscious influences on unconscious perception. *Journal of Memory and Language*, 28, 412-443.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy, 60,* 685-700.

Debner, J. A. & Jacoby, L. L. (1994). Unconscious perception: Attention, awareness, & control. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 20, 304-317.

Dennett, D. C. (1978). Brainstorms. Montgomery, VT: Bradford.

Diamond, A., & Goldman-Rakic, P. S. (1989). Comparison of human infants and infant rhesus monkeys on Piaget's AB task: Evidence for dependence on dorsolateral prefrontal cortex. Experimental Brain Research, 74, 24-40.

Dienes, Z. (1992). Connectionist and memory array models of artificial grammar learning. Cognitive Science, 16, 41-79.

Dienes, Z., & Altmann, G. (1997). Transfer of implicit knowledge across domains? How implicit and how abstract? In D. Berry (Ed.), *How implicit is implicit learning?* (pp 107-123). Oxford: Oxford University Press.

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin and Review, 4,* 3-23.

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., 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., Kurz, A., Bernhaupt, R., & Perner, J. (1997). Application of implicit knowledge: deterministic or probabilistic? Psychologica Belgica, 37, 89-112.

Dokic, J. (1997). Two metarepresentational theories of episodic memory. Paper presented at the Annual Meeting of the ESPP in Padua, Italy August 1997. (unpublished)

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes.* Cambridge, MA: MIT Press.

Dretske, F. (1995). *Naturalizing the mind.* Cambridge (Massachusetts), London: The MIT Press.

Dulany, D. E. (1991). Conscious representation and thought systems. In R.S. Wyer & T.K. Srull (Eds), *Advances in social cognition, vol 4* (pp. 97-120). Erlbaum: Hillsdale, NJ.

Dulany, D. E. (1996). Consciousness in the explicit (deliberative) and implicit (evocative). In J. D. Cohen & J. W. Schooler (Eds), *Scientific approaches to the study of consciousness* (pp 179-212). Erlbaum: Hillsdale, NJ.

Eriksen, C. W. (1960). Discrimination and learning without awareness: A methodological survey and evaluation. *Psychological Review, 67*, 279-300.

Evans, G. (1975). Identity and predication. *The Journal of Philosophy*, 72(13), 343-363.

Field, H. (1978). Mental representation. *Erkenntnis, 13,* 9-61.

Fodor, J. A. (1983). *The modularity of mind*. Canbridge, Mas.: MIT Press.

Fodor, J. A. (1987). A situated grandmother? Some remarks on proposals by Barwise and Perry. *Mind & Language*, 2, 64-81.

Fodor, J.A. (1978). Propositional attitudes. *The Monist, 61*, 501-523.

Fodor, J.A. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In J.L. Garfield (Ed.), *Modularity in knowledge representation and natural-language understanding.* (pp. 25-36). Cambridge (Massachusetts), London: The MIT Press.

Fowler, C. A., Wolford, G., Slade, R. & Tassinary, L. (1981). Lexical access with and without awareness. *Journal of Experimental Psychology: General*, 110. 341-362.

Gardiner, J. (1988). Functional aspects of recollective experience. *Memory and Cognition,* **16**, 309-313.

Gentilucci, M., Chieffi, S. & Daprati, E. (in press). Visual illusion and action. *Neuropsychologia*.

Gewei; Y., & van-Raaij, F. W. (1997). What inhibits the mere-exposure effect: Recollection or familiarity? *Journal of Economic Psychology, 18*, 629-648.

Gibson, J. J. (1950). *The perception of the visual world.* Boston: Houghton Mifflin.

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review*, **100**, 279-297.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16,* 1-113.

Gordon, R.M. (1995). Simulation without introspection or inference from me to you. In M.Davies & T.Stone (Eds.), *Mental Simulation: Evaluations and applications.* (pp. 53-67). Oxford: Blackwell.

Greenwald, A.G. (1992). New look 3: Unconscious cognition reclaimed. *American Psychologist, 47*, (6). 766-779.

Güzeldere, G. (1995). Is consciousness the perception of what passes in one's own mind? In Th.

Metzinger (Ed.), Conscious experience (pp. 335-357). Paderborn: Schöningh.

Heyes, C. & Dickinson, A. (1993). The intentionality of animal action. In M. Davies & G.W. Humphreys (Eds.), *Consciousness* (105-120). Oxford: Blackwell.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *The Behavioral and Brain Sciences*, 9, 1-66.

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

Jacoby, L.L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, 110, 306-340.

Jacoby, L.L., Lindsay, D.S., & Toth, J.P. (1992). Unconscious influences revealed: Attention, awareness, and control. *American Psychologist, 47*, 802-809.

Karmiloff-Smith, A. (1986). From meta-processes to conscious access: Evidence from children's metalinguistic and repair data. *Cognition*, **23**, 95-147.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.

Kihlstrom, J.F. (1996). Perception without awareness of what is perceived, learning without awarenes of what is learned. In M. Velmans (Ed.), *The science of consciousness: Psychological, neuropsychological and clinical reviews* (pp 23-46). London, New York: Routledge.

Kihlstrom, J., Barnhardt, T., & Tataryn, D. (1992). Implicit perception. In R. Bornstein & T. Pittman (Eds.), *Perception without awareness: cognitive, clinical, and social perspectives.* London: Guilford Press.

Kirsh, D. (1991). When is information explicitly represented? In P. Hanson (Ed.). *Information, thought, and content.* UBC Press.

Künne, W. (1995). Some varieties of thinking. Reflections on Meinong and Fodor. *Grazer Philosophische Studien*, **50**, xxxxxxx.

Leslie, A. (1994) Pretending and believing: Issues in the theory of ToMM, *Cognition, 50*, 211-238.

Leslie, A. M. (1987). Pretense and representation: The origins of "Theory of Mind." *Psychological Review,* **94**, 412 426.

Lewis, D. (1986). Causal explanation. In D. Lewis (Ed.), *Philosophical papers (Vol. 2).* Oxford University Press.

Manza, L., & Reber, A. S. (1997). Representation of tacit knowledge: Transfer across stimulus forms and modalities. In D. Berry (Ed.), *How implicit is implicit learning?* (pp 703-106). Oxford: Oxford University Press.

Marcel, A. J. (1983a). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, 15, 197-237.

Marcel, A. J. (1983b). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology*, 15, 238-300.

Marcel, A. J. (1993). Slippage in the unity of consciousness. In*: Experimental and theoretical studies of consciousness* (Ciba Foundation Symposium 174, 168-186). Chichester: Wiley.

McCarthy, J., and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Mehler and D. Michie (Eds.), *Machine intelligence,* Vol 4. Edinburgh: Edinburgh University Press.

Merikle, P. M. (1992). Perception without awareness: Critical issues. *American Psychologist, 47,* 792-795.

Millikan, R. G. (1984). *Language, thought, and other biological categories.* Cambridge, MA: MIT Press.

Milner, D. A. & Goodale, M. A. (1995). Visual pathways to perception and action. In T. P. Hicks, S. Molotchnikoff. & Y. Ono (Eds.), *Progress in Brain Research, vol. 95* (317-337). Elsevier Science Publishers.

Nichols, S. and Stich, S. (1998). *Pretense and Counterfactuals: Possible Worlds in Cognitive Science.* (unpublished).

Norman, D.A. & Shallice, T. (1980). Attention to action: Willed and automatic control of behaviour. Center for Human Information Processing Technical Report No. 99. Reprinted in revised form in *Consciousness and self regulation, Vol. 4* (ed. by R.J. Davidson, G.E. Schwartz & D. Shapiro, pp. 1-18). New York: Plenum 1986.

Paillard, J., Michel, F., & Stelmach, G. (1983). Localization without content. A tactile analogue of 'Blindsight'. *Archives of Neurology,* 40, 548-551.

Perner, J. (1990). Experiential awareness and children's episodic memory. In W. Schneider and F. E. Weinert (Eds.), *Interactions among aptitudes, strategies, and knowledge in cognitive performance* (pp. 3-11). New York, Berlin, Heidelberg: Springer Verlag.

Perner, J. (1991). *Understanding the representational mind.* Cambridge, MA: Bradford Books/MIT-Press.

Perner, J. (1998). The meta-intentional nature of executive functions and theory of mind. In P. Carruthers & J. Boucher (Eds.), *Language and thought* (pp.270-283). Cambridge: Cambridge University Press.

Perner, J. & Clements, W. A. (in press). From an implicit to an explicit theory of mind. In Y. Rossetti & A. Revonsuo (Eds.), *Interaction between dissociated implicit and explicit processing*. Amsterdam: John Benjamins.

Perner, J., Leekam, S.R., & Wimmer, H. (1987). Three-year olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology, 5,* 125-137.

Perruchet, P., & Gallego, J. (1997). A subjective unit formation account of implicit learning. In D. Berry (Ed.), *How implicit is implicit learning?* (pp 124-161). Oxford: Oxford University Press.

Perry, J. (1986). Thought without representation. *Supplementary Proceedings of the Aristotelian Society*, 60, 137-166.

Piaget, J., & Inhelder, B. (1941/1974). *The child's construction of quantities: Conservation and atomism.* (A. J. Pomerans, transl.) New York: Basic Books.

Pöppel, E., Held, R., & Frost, D. (1973). Residual visual function after brain wounds involving the central visual pathways in man. *Nature*, 243, 295-296.

Reason, J.T., & Mycielska, K. (1982). *Absent minded? The psychology of mental lapses and everyday errors.* Englewood Cliffs, NJ: Prentice Hall.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behaviour*, *6*, 855-863.

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

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

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

Reingold, E. M., & Merikle, P. M. (1993). Theory and measurement in the study of unconscious processes. In M. Davies & G. W. Humphreys (Eds.), *Consciousness* (40-57). Oxford: Blackwell.

Richardson-Klavehn, A. & Bjork, R. A. (1988). Measures of memory. Annual Review of Psychology, 39, 475-543.

Richardson-Klavehn, A., Gardiner, J. M., & Java. R. I. (1994). Involuntary conscious memory and the method of opposition. *Memory*, 2, 1-29.

Richardson-Klavehn, A., Gardiner, J. M., & Java, R. I. (1996). Memory: task dissociations, process dissociations, and dissociations of consciousness. In G. Underwood (Ed.), *Implicit cognition* (pp.85-158). Oxford: Oxford University Press.

Roberts, P. L., & McLeod, C. (1995) Representational consequences of two modes of learning. *Quarterly Journal of Experimental Psychology, 48A*, 296-319.

Roediger, H. L. III, & McDermott, K. B. (1996). Implicit memory tests measure incidental retrieval. Paper presented at the XXVI International Congress of Psychology, Montreal, August, 1996.

Rosenthal, D.M. (1986). Two concepts of consciousness. *Philosophical Studies*, **49**, 329-359.

Rossetti, Y. (1998). Implicit short-lived motor representation of space in brain-damaged and healthy subjects. *Consciousness & Cognition, 7,* **(in Ref. Manag. noch ”in press”)**

Russell, B. (1919). On propositions: What they are and what they mean. *Proceedings of the Aristotelian Society, 2:*1-43.

Salmon, W.C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Schacter, D. L. (1987). implicit memory: History and current status. *Journal of Experimental Psychology: Learning, memory and cognition*, **13**, 501-518.

Schacter, D. L., Bowers, J., & Booker, J. (1989). Intention, awareness, and implicit memory: The retrieval intentionality criterion. In S. Lewandowsky, J. C. Dunn, & K. Kirsner (Eds.), *Implicit memory: Theoretical issues* (pp. 47-65). Hillsdale, NJ: Erlbaum.

Searle, J. (1983). *Intentionality.* Cambridge: Cambridge University Press.

Shallice, T. (1988). Specialisation within the semantic system. Special Issue: The cognitive neuropsychology of visual and semantic processing of concepts. *Cognitive Neuropsychology*, *5,* 133-142.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge University Press: Cambridge.

Shanks, D. R. & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioural and Brain Sciences*, 17, 367-448.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3-22.

Smith, N. & Tsimpli, I-A. (1995). *The mind of a savant: Language-learning and modularity.* Oxford: Blackwell.

Sperber, D. (1996). *Explaining culture: A naturalistic approach.* Oxford: Blackwell.

Sperber, D. (1997). Intuitive and reflective beliefs. *Mind & Language,* **12**, 67-83.

Squire, L.R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* (2). 195-231.

Stadler, M. A., & Frensch, P. A. (Eds) (1998). *Handbook of Implicit Learning*. Thousand Oaks, USA: Sage.

Strawson, P. F. (1959). *Individuals*. London: Methuen.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology,* **26**, 1 12.

Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind.* Cambridge (Massachusetts), London: The MIT Press.

Ungerleider, L. & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of motor behavior* (pp. 549-586). Cambridge, MA: MIT Press.

Weiskrantz, L. (1988). Some contributions of neuropsychology of vision and memory to the problem of consciousness. In A. J. Marcel & E. Bisiach (Eds.), *Consciousness in contemporary science* (pp. 183-199). Oxford: Clarendon Press.

Weiskrantz, L., Warrington, E. K., Sanders, M. D., & Marshall, J. (1974). Visual capacity in hemianopic field following a restricted occipital ablation. *Brain,* 97, 709-728.

Whittlesea, B. W. A., & 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.

Winograd, T. (1975). Frame representations and the declarative-procedural controversy. In D. G. Bobrow & A. Collins (Eds.*), Representation and understanding* (pp. 185-210). Studies in cognitive science. New York: Academic Press.

Wong, E. & Mack, A. (1981). Saccadic programming and perceived location. *Acta Psychologica*, 48, 123-131.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monographs*, 9(2, pt. 2), 1-27.

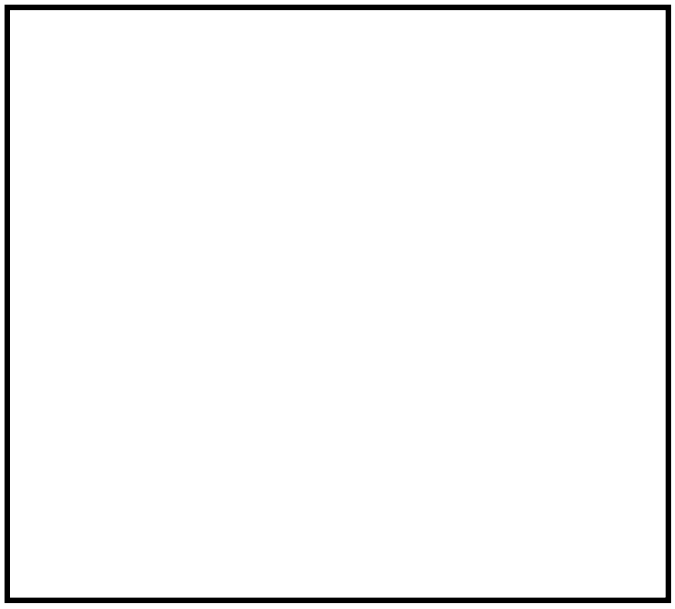
Zelazo, P.D., Reznick, J.S., & Pinon, D.E. (1995). Response control and the execution of verbal rules. *Developmental Psychology, 31,* 508-517.

Table 1.
Possible Combinations of Implicit & Explicit Knowledge of Aspects of Facts.
(Factuality stands for factuality and/or time).

	represented	
	explicitly	implicitly
1.	property	individual + predication + factuality
2. (a)	property + individual	predication + factuality
(b)	property + predication	individual + factuality
3.	property + individual + predic.	factuality
4.	property + ... + factuality	none

Table 2.

Laid down representation of fact that Fb		Memorial state of awareness	Retrieval volition	Reference by:	Recognition test response
Property	"F"	none	involuntary	nothing	correct guess.
Compound	”F-X”	feel of famil.	--”--nothing		recogn. by famil.
Predication	"Fb"	--"--.	direct vol.	"part of list"	--"--
Factuality					
+Time	"Fb happened"	knowing past	--"--	"was on list"	"K" (past event)
Origin	"I experienced Fb"	remembering	--"--	"remember!"	"R"



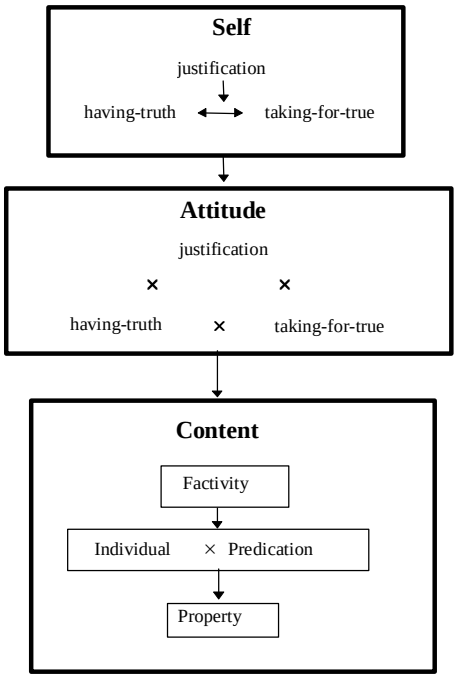


Figure 1.
Constraints on explicitness. An arrow denotes that explicitness of the item from which the arrow emanates entails explicitness of the item to which the arrow points. An "×" denotes that the explicitness of the two terms can be varied freely.