Authors' Response

Deconstructing RTK: How to Explicate a Theory of Implicit Knowledge

Josef Perner, University of Salzburg, Austria

Zoltan Dienes, University of Sussex, UK

7 April 1999

Acknowledgements.

We thank Ingar Brinck and Ron Chrisley for guidance through the mysteries of nonconceptual content, Ron Chrisley for other useful discussions, and Bruce Bridgeman for making his data available to us.

Abstract

In this response, we start from first principles building up our theory to show more precisely what assumptions we do and do not make about the representational nature of implicit and explicit knowledge (in contrast to the target article, where we started our exposition with a description of a fully fledged RTK). Along the way, we indicate how our analysis does not rely on linguistic representations but it does imply that implicit knowledge is causally efficacious; we discuss the relationship between property structure implicitness and conceptual and nonconceptual content; then we consider the factual, fictional and functional uses of representations, and how we go from there to consciousness. Having shown how the basic theory deals with foundational criticisms, we indicate how the theory can elucidate issues commentators raised in the particular application areas of explicitation, voluntary control, visual perception, memory, development (with discussion on infancy, TOM and executive control, and gestures), and finally models of learning.

1. Deconstructing RTK (the representational theory of knowledge).

Several commentators have criticised us on points that seem to be consequences of our adopting RTK (the Representational Theory of Knowledge) as a framework for our exposition. In fact we had not explicitly used RTK (or RTM) in our original draft. We introduced it in a revision with the aim to provide readers with a familiar framework detailing the elements of propositional attitudes, without wanting to buy into the usual interpretation of being language like (Fodor, 1975: "A language of thought"). In other words, our strategy (as it finally appeared) was top-down to start with the most explicit, the most elaborate human understanding of knowledge and then decompose it into its elements. Since the starting point is highly permeated by language this created the wrong impression of what we are trying to achieve. So, we take to heart **Gall**'s admonition that we rely too heavily and too early on RTK and **Carlson**'s advice to invert our focus. So, we now try to trace our enterprise in the opposite direction from the bottom up. This may help to allay some of the fears about the core assumptions underlying our analysis. To overview the issues, whether we start top down or bottom up, we do presume that having knowledge or holding a belief does involve explicit representation, and to that degree we do hold a representational theory of knowledge. A fully fledged RTK holds that one can know a proposition *p* only if *p* is itself explicitly represented. Thus, for fully explicit knowledge, we hold that RTK is strictly true. For implicit knowledge, we also hold there must be explicit tokening of some representation. But in contrast to RTK, we do allow implicit knowledge that does not consist in a representation tokening the full proposition *p*. It is only to that degree that our framework is not a RTK. Our commentators must bear in mind that subscribing to these assumptions does not entail subscribing to all

other assumptions of a Fodorian world view; for example, the assumptions of RTK that we do hold can be (perhaps should be) held even by a rabid connectionist, a point we return to in Section 7 below.

1.1 Overly linguistic.

Several commentators complained that our analysis of implicit-explicit knowledge is overly linguistic (**Carstairs-McCarthy, Pietroski & Dwyer, Jimenez & Cleeremans**) and anthropomorphic (**Mercado & Murray**) because we are relying heavily on the representational theory of mind (RTM) or knowledge (RTK). It is true that the analysis starts from an analysis of ordinary language expressions about the mind (that's what philosophers of mind are mainly engaged in). But this starting point is hard to avoid. Even behaviourists usually rely on anthropomorphic descriptions of what the animal is doing: pressing a lever, jumping through a hoop. Of course, as research progresses it moves away from that starting point and develops better analyses for the specific matter of investigation. However, on occasion it is important to remind oneself of the starting point because dedicated research often forgets some useful distinctions. For instance, memory research for many years had lost the distinctions that are re-evoked in the implicit-explicit and in the semantic-episodic distinction (Tulving, 1985) and that were originally referenced by the old masters, e.g., Ebbinghaus (1885) and James (1890).

Evidently, these distinctions have been made primarily on the basis of our linguistic distinctions and our phenomenology. However, there is no reason why these distinctions could not be separated from their linguistic and introspective origins in order to investigate the presence of these processes in non-linguistic animals. One example of such an enterprise is the work by Dickinson (e.g., Hayes & Dickinson, 1993) who asked whether rats do or do not represent propositional attitudes, like their goals and intentions. In fact, one purpose of our analysis is to conceptually penetrate the reason why in the human case language use,

consciousness, voluntary control, directness of tests, etc. (a point appreciated by **Evans & Over**) tend to go together in order to be able to design experiments that do not rely on linguistic competence.

1.2 Causally inert.

At the base level we are concerned about representations. As a best shot at a quick definition: representation are states (typically internal) of an organism that are about something (typically the organism's environment). They get their aboutness by the fact that they causally govern the organism's interaction with its environment by mapping the relevant distinctions in the environment. And they can only map the environment non-accidentally if there is a causal process from environment to representation (e.g., perception). Environmental differences that are reflected (encoded) in the representation are represented explicitly. Now, the interesting thing here is, that even if my representational capacities only allow me to make a difference between lion and (domestic) cat which then controls relevant behaviour, if the cat in front of me makes my mind go into its cat state, then that state represents *implicitly* that there is a cat and not just cat-ness.

Even though it is as implicit as they come, this representation is NOT causally inert as **Jimenez and Cleeremans,** but also **Carlson** and **Vokey and Higham** (latent knowledge—completely without effect) suggest. What one could say, is that implicit knowledge has fewer causal effects than more explicit knowledge, since more explicit knowledge allows more internal distinctions which can lead to a greater variety of causal effects. But implicit knowledge is not causally inert! In defence of these claims of causal inertness, one could surmise that these commentators interpreted ”implicit knowledge” as referring only to that aspect of implicit knowledge which remains implicit. Since these aspects are not reflected in internal differences they can not have any causal consequences on behaviour—so their likely

reasoning. However, even that is not quite right. The reason why the implicit aspects are ‘represented’ at all is because they are involved in the causation of the internal representation: if it weren’t for the fact that it was the particular cat that is responsible for my mental ”cat” token then the fact that it was that particular individual which is a cat, would not be implicit in my explicit ”cat”. Moreover the causal role of the implicitly represented individual is also important for the appropriateness of my behavioural effects of the explicit parts, e.g., saying ”cat” and smiling as opposed to saying ”lion” and running away in fright. If there weren’t a particular individual or if it weren’t for real then my behaviour would be inappropriate. What the implicit-explicit distinction captures is where the causal effects are located: in the environmental setting (implicit) or in the internal distinctions (explicit). It, thus, captures an important aspect of the substance matter and is not just a theory of how scientists use the terms as **Taatgen** suggests.

The Implicit Piggybacks on the Explicit. At this point one may also wonder whether **O’Brien and Opie**’s characterisation of ”the implicit piggy backing on the explicit” is a fully accurate characterisation of our position. One interpretation of this characterisation is that implicit aspects depend counterfactually on explicit aspects. True, if there were no explicit representation of lion vs. cat then there would be no aspects implicit in anything. However, if the source of the implicit aspects, that is, the fact that the particular individual is the cause of the explicit distinctions, were not existent then there would be no explicit distinction. So the explicit is also piggybacking on the implicit.

Explicit Individuals with Implicit properties. The causal role that properties and individuals play in knowledge formation provides a good context for addressing the question whether there is an implicit-explicit hierarchy such that properties can be explicit with

individuals as the carriers of that property remaining implicit but not the other way around. As **Barber** correctly observed, the main purpose of our analysis is to lay open the possible elements according to which knowledge can be implicit or explicit. Nevertheless, we also had the intuition that not all combinations are possible and tried to formulate a partial hierarchy. Barber agrees to our intuition that there is some asymmetry but challenges our specific proposal with a counterexample. In his variant naming game the player is confronted with several individuals of whom one is being highlighted at each turn. The player just identifies explicitly (mentally as well as verbally) the particular individual but makes no internal distinction concerning the property of being highlighted. The player relies implicitly on the fact that his identification is being taken to refer to what ever is being highlighted.

We agree that this is an intriguing counterexample and our answer is not one of the two unviable options anticipated by Barber. Rather we want to point out that the counterexample seems to work only with specific properties like being highlighted which is not primarily a property of an object but a property that describes the interaction between the object and the players of the naming game. The lesson we take from this observation is that, whether something can be left implicit or depends on the causal relationship between these aspects and the observer (game player). Because the highlighting causes the player to attend to the particular individual the property of being highlighted can be left implicit. In general, however, there remains an asymmetry between individuals and properties: it is necessary that some property be represented explicitly, namely the one that individuates the object in the observer's mind, before any individual can be identified.

Westerberg and Marsolek deny that either the individual or the property can remain implicit, because in the naming game the player has to represent that "cat" applies to that particular individual, and in the subliminal Stroop experiments one has to represent which colour word was presented in that trial. Our point, however, is that if one doesn't predicate the

perceived properties to any particular event or individual but simply answers with whatever colour comes to mind first, one can be above chance correct because the most recently presented colour word makes it more likely that it comes to mind first. The subject then makes an inference, predicating the colour to a particular trial, but this inference occurs some time after the moment of perception itself.

In fact, the physiological evidence, mentioned by **Westerberg and Marsolek,** that visual properties are encoded separately and later bound together (predicated to a single individual) illustrates the possibility that on occasion only the property information could make it into higher brain regions without the binding information. This could still have some behavioural effect, whereas if the property information goes lost and only the binding marker survives then it is hard to imagine what behavioural effect this could have. In early vision, location is initially coded property-structure implicitly in spatiotopic feature maps where there is not a single representation for a particular location, but a lot of location-feature representations. Hence the individual is not coded explicitly, but binding to an individual object is still possible at a later stage of processing as the result reported by **Bridgeman** suggests.

1.3 Property-Structure, Predication and Non-conceptual Content (NCC).

Brinck asks how nonconceptual content (NCC) fits into our framework. NCC bears an interesting relation to property-structure implicitness. Chrisley (1996) defined NCC as content not entirely composed of constituents that meet the generality constraint (i.e. the constraint that constituents can freely recombine with each other). A representation that carries NCC with constituent structure would thus be property structure implicit. Say the nonconceptual content in question is *green and small*, which is NCC if the constituents don't satisfy the generality constraint. So *green and small* is not represented by an all-purpose *green* token

concatenated with an all-purpose *small* token. Thus, the structure of being green and being small is not made explicit by the representation for green and small; it is property structure implicit. Further, the representational content meets the definition of NCC given by Brinck because the holder of this content can have it without having the concepts *green*, *small*, etc. that we use to describe the content.

On Cussins' (1992) view NCC cannot be predicated of an external (conceptually identified) object (since NCC does not necessarily respect the boundaries of such objects). Consequently, NCC cannot have a truth value because only expressions that predicate properties of individuals have a truth value (Evans, 1975) in the classical sense of being able to derive contradictions. This view conforms with **Brinck**'s characterisation of NCC as having correctness conditions without being able to have a truth-value assigned. When this position is combined with the claim that NCC is accessible to consciousness and volitional control it poses a problem for our claim that explicit predication is prerequisite for consciousness and volitional control.

However, several theorists take a different view. Chrisley (1996), Peacock (1993) and Bermudez (1995) do regard NCC as propositional and capable of having a truth value (there is a way of predicating that allows this). Thus, on these views NCC poses no problem for our framework: It may be represented maximally implicitly as a property, or fully explicitly, as a representation of knowing an individual has a certain property (conscious but not verbalisable because the property cannot be conceptualised).

NCC interpreted as structure-implicit representations makes also clear that our immediate action regulation is based on NCC. For, our interaction with the world involves representations that structure-implicitly represent a mix of object properties and of features of how to act on these objects, since this is the most efficient way of effecting action (e.g., common coding of perception and action, Prinz, 1990). Normal action execution is, therefore,

difficult to verbalise, as NCC cannot be dissected with our concepts. However, under the assumption that NCC is predicable, this allows, as **Brinck** observes, for the intentional and wilful improvement of craftsmanship through perceptual monitoring in the absence of verbal reflection.

Why should verbal reflection come into the picture? We suspect because the mention of predication and propositional conjures up images of 'language-like representation' as in the analogue vs. propositional representations dispute (Pylyshyn, 1973; Kosslyn, 1975). There is of course some link between the propositional and the linguistic. Linguistic expressions are characterised by a high degree of articulation of their meaningful parts, i.e., basic units of meaning (words) are linked by precise rules of concatenation into larger meaningful units (sentences). Images as prototypical analogue representations are meaningful without having any clearly separable parts. In order to have an explicit representation of predication a minimal degree of articulation is needed for linking the predicate to its subject. However, no further degree of articulation is needed for the predicate. It could be an image. In any case, predication in this view is something very fundamental and not just a feature of language as **Carstairs-McCarthy** puts forward and it is something that animals must be capable of if they engage in variable binding regardless of their linguistic abilities.

For example, summaries of various features of NCC (Brinck, 1997; Peacock, 1993) list the finer grain of visual images as one feature of NCC. Like the detailed imagistic schema of faces by which we are able to recognise so many different people, the content of images consists of properties that can be predicated to objects or events in the world. And because their content can be predicated, this predication and its factuality and eventually our knowledge thereof can be made explicit and they can be consciously experienced, even if they cannot be completely described.

1.4 Concepts and Property Structure.

According to our assumptions, one possesses a concept of a property only if one has the internal distinction whose function it is to indicate that property. It is purely defined by its semantic/symbolic relation to the world. Hence it is a distinction which is predicable to the particulars in the world that carry the property. However—and Fodor couldn't object—these conceptual distinctions can only fulfil their semantic function if they are embedded in processes among which other distinctions are made, many of them being non-conceptual and property-structure implicit in a way that cannot be explicated. Since it is implicit it cannot be coherently addressed for different purposes which may explain why people give idiosyncratic responses when questioned about it and produce incompatible results for different tasks like rank ordering definitional properties as opposed to rank ordering category instances by typicality, as observed by **Hampton**. Conceptually defined criterial properties may play little role in typicality judgements.

In this context **Hampton** raises the difficult question about what properties are structure implicit in other properties. Hampton suggests that all contingent implications, like being composed of cells containing DNA is property-structure implicit in bachelor. This seems to go too far, violating the linguistic intuition of what is conveyed implicitly when I say "he is a bachelor." With this I do not convey implicitly that he is made up of DNA carrying cells. To avoid this consequence, we formulated our criterion in terms of meaning. Not just any supporting fact makes for implicitness but only the ones "that are necessary for the explicit part to have the meaning it has." (Section 1, penultimate paragraph).[1]

There used to be the distinction between analytic and synthetic truths, which has got bad press since Quine's (1951) paper. However, the intuition behind it does not go easily away.

[1] A fuller exposition would make clear that the structure implicit facts are those that define the conditions that must hold only in *nearby* possible worlds for the representation to have the meaning it has. Facts like laws of physics, chemistry, or biology must hold even in relatively distant possible worlds; if they did not hold, the

Keil (1989) has made good use of a distinction between definitional and characteristic features in children's acquisition of concepts. This distinction underlies the strong intuition that sensitivity to some features is essential for a proper understanding of a concept while others aren't. As far as we can tell the question of how to make this distinction is an unsolved issue. We can only rely on our natural linguistic intuition.

One interesting question is what role this distinction between defining and characteristic features may play in the "externalist" view on concepts shared by Fodor, that the concept is purely determined by its semantic relations to a property. One possibility (Keil, 1998; Perner, 1998) is that defining properties need to be internally distinguished in order that the target distinction can serve its representational function. In that case the concept bachelor would necessitate conceptual or nonconceptual sensitivity to maleness and being unmarried but no such sensitivity for detecting the presence of DNA, cells, etc. Moreover, despite the required sensitivity to maleness and being married no definitions in terms of the corresponding concepts need to be formed.

Also logical implications are not necessary for meaning. A mathematician who knows Peano's axioms does, thereby, not implicitly know all of mathematics that is entailed by them. So, on our definition of property-structure implicit the case in Plato's Meno where the young boy is led by his teacher to draw out the implications of what he already knows is, in agreement with **Homer and Ramsay,** not a case of making property-structure implicit knowledge explicit. Contrary to Homer and Ramsay, conscious reflection is sometimes not sufficient for making property-structure implicit knowledge explicit, as we discussed in the case of NCC. To give an empirical example (of, as it were, NCC relative to a specific domain

---

world, and not just the meaning of a few representations, would be completely different. (Thanks to Ron Chrisley

and task), Roberts and McCleod (1995) found that people trained under full attention to recognise exemplars of the category, e.g., ”triangle and red” were equally good with a monochrome display which only showed shapes without colours at indicating triangles as *possible* instances of the category, but were rather poor at recognising the triangle as a possible instance after learning with diverted attention.

Contrary to **Overskeid**'s claim that representing a compound property ipso facto explicitly represents its components, the Roberts and McCleod paper shows that property structure implicitness is not only logically possible but empirically observable. (This can be achieved by representing the components in a context sensitive way; i.e. their only function is to indicate the component when the other components are present; thus, each component is not explicitly represented in itself.)

1.5 Factual, Fictional, and the Functional Use of Representations.

At the bottom a strict separation of representation and functional use is not possible since (by our quick-shot definition) representations do not just map the environment but also govern the interaction with that environment. A relative separation of representation from their use emerges in more complex systems as the articulation of components increases. Imagine a connectionist robot that can learn to negotiate an environment to get to a particular goal. The representation of the goal may be enmeshed with the representation of the given environment because when the goal changes the robot has to relearn much about the layout of the environment. In this case, there is no systematic internal distinction of the two basic functional uses: beliefs and desires. This distinction is property-structure implicit in a representation compounding information about the environment and where the robot wants to be. For a system that can flexibly combine knowledge of the environment with its goal this separation

---

for making this suggestion.)

needs to be made by some internal distinction. Presumably, goal devaluation studies show that rats can make this distinction (Hayes & Dickinson, 1993). Propositional attitudes, even though they come from a 'linguistic' analysis, can be studied in non-linguistic animals, pace **Mercado and Murray** and **Carstairs-McCarthy**.

The necessary internal distinction can be implemented in many different ways. In the philosopher's favourite metaphor of a belief and desire box, or as functional markers on individual representations. Its prime purpose is to ensure the proper use of the thus marked representations. However, since by doing so it also classifies the marked representations as representing the environment or the organism's goal these functional markers (for beliefs and desires) also qualify as representations constituting procedural knowledge of the distinction between facts and goals. It is not declarative knowledge. For it to become declarative the distinction has to come under the scope of the belief marker. Only then does one know (believe) that something is a fact.

In the target article we were not concerned much about the belief-desire distinction but about the further distinction between factuality and fiction. The problem can easily be seen. With only a belief-desire distinction I can only know (believe) or want something. I can't just think of something. Well, there is one degenerate possibility: unpredicated properties. Because they are unpredicated they do not describe a fact, hence they remain non-factual (but they are not exactly fiction, either). The question of how to introduce the factual-fictional distinction properly has recently been discussed by Nichols & Stich (1998 manuscript) and Currie and Ravenscroft (in press, Ch. 5). Nichols and Stich suggest to introduce a third box, namely a PW-box, a possible-world box (or type of functional marker—perhaps the omission of one of the other two). With this we gain a functional distinction between factuality and fiction.

We agree that such a functional distinction is at the heart of the factual-fictional distinction (and all hypothetical reasoning as **Evans and Over** point out), as it is for the fact-

goal distinction. Hence we agree with **Currie** that we can't capture the factual or fictional status purely within the content, it can only be captured by the functional distinction. However, as our bottom-up analysis—pursued here—shows, making the functional distinction implies a representational distinction, i.e., the functional marker makes the distinction explicit, though only as a property without explicit predication. This amounts to predication-implicit procedural knowledge of the distinction. In other words, making factuality explicit means introducing a functional (representational) distinction that has the appropriate effects. **Currie**'s question only arose because our analysis pursued a top-down analysis with RTK as a starting point.

The bottom-up analysis also raises another interesting question not apparent in our original treatment. The question is whether a purely functional distinction providing procedural knowledge of the factual-fictional distinction is sufficient. This question has recently been put into focus by Nichols and Stich (1998) in a discussion of pretend play. That is, when pretending that this (banana) is a telephone the infant simply switches to a different functional mode concerning the representation, "the banana is a telephone," without knowing that she is pretending (since the functional use is not registered within the belief box).

Although this is perfectly possible, the intuition among developmental psychologists (e.g., Leslie, 1987; Piaget, 1945) is that pretence emerges with the knowledge that one is pretending (in some minimal sense). Piaget spoke of the infant's "knowing smile" as an indicator of this reflective awareness. Moreover, Nichols and Stich's suggestion puts hypothetical reasoning including pretence on a par with the belief-desire distinction. It follows that one's pretence should be able to remain as unconscious as our desires. However, though in our many automatic actions we are often not aware of the reasons for why we are doing what we are doing, the same cannot be said for pretence.

Like the developmental intuition, our phenomenal self insight likewise suggests that pretence (and hypothetical reasoning, etc.) does not occur unconsciously. The fact that we are pretending is always within our belief box. Perner (1991, chapter 2)—following Leslie's (1987) analysis of pretence—suggested that the real-hypothetical (i.e., factual-fictional) distinction is based on meta-representational context markers which serve a functional and representational role (see also Sperber, 1997 for a similar suggestion). In our current terminology we can say that the factual-fictional distinction does not emerge first as procedural knowledge, but comes directly as declarative knowledge.

Pursuing this option, that the factual-fictional distinction consists of a functional distinction within the belief box (or within the scope of the fact marker distinguishing facts from goals) we can answer another critic. An organism that only distinguishes functionally between facts and goals (belief and desire box) cannot represent the fictional vis à vis the factual. For such an organism the quisitive suggestion by **Nichols and Uller** to have a standard rule: if p is believed then one can add "It is a fact that p", is perfectly possible but useless, since every occurrence of _ in the belief box has the function of being taken as a fact. The dorsal action system may be of this kind, provided it processes propositions at all. The rule, however, fails when it becomes relevant, namely in an organism (or our ventral visual information processing path) that can distinguish between factual and fictional with appropriate functional markers. If that organism encounters some proposition p without a marker, then the rule would be dangerous to apply. The claim is that such propositions can float around in our head (belief box). They constitute implicit knowledge of the fact that p, because they have been properly caused by perceiving p, but their factuality has not been explicitly marked. So they remain factuality-implicit knowledge.

Making the factuality-fiction distinction dependent on markers within the belief box makes it akin to the distinction between temporal contexts: knowing what happened now and

knowing what happened earlier. There is varied support that the ability to distinguish fact from fiction and representing temporal contexts go hand in hand. As **Boucher** points out explicit representation of time is part of a cluster of abilities for which it is controversial whether animals have them and that autistic children have difficulty with. And there is also evidence from normal development that the ability to pretend emerges at the time children can represent earlier states of affair to understand invisible displacement of objects (Perner, 1991). Our disagreement with Boucher is that it is not clear to us whether explicit representation of time is the driving force behind these new abilities rather than the more general ability to differentiate contexts within the belief box. We also have difficulty seeing how the development of time keeping mechanisms provides an explicit representation of temporal contexts.

1.6 From Predication and Factuality to Consciousness.

As we introduced RTK into the revision of the target article we also cut down on the issue of how to define implicit and explicit knowledge. The relevant passage in the original submission clarified how the implicit-explicit distinction defined for linguistic expressions and representations applies to knowledge, a transition that **Gall** found wanting.

> ”Knowledge of a fact or an aspect of a fact is *explicit*, if that fact or aspect is represented by an internal state whose function it is to covary with it. Other, supporting facts or aspects of facts, that are not explicitly known but which must hold, in order for the explicitly known fact to be known, are *implicitly known*.” (original draft of target article).

We refined the notion of knowledge by specifying 4 conditions (2.1.2). **Smith** objects to these conditions because in his view they conflate two standard accounts of knowledge. Indeed we did not spell out the relationship between representation and content in any detail, we only

indicated it. For instance, the formulation of ”(i) R is accurate (true)” the parenthesis is to indicate that ”the proposition represented by R is true,” just as Smith suggests. Our four conditions specify primarily the causal account and the information in parenthesis or in subordinate clauses indicate how the particular causal condition relates to logical/foundational aspects. We can't see why this should be so objectionable. Our approach far from conflating two theories of knowledge appropriately allows the person to believe either theory of knowledge. In any case, however we specify conditions of knowledge, it is difficult to see how that would invalidate what we have to say about the implicit-explicit distinction.

We also argued that making the attitude of knowing explicit requires that the content be made explicit (2.1.3), in particular that it requires explicit factuality and predication. Several commentators suggested counterexamples to this claim. **Bibby and Underwood** point out that one can represent ”I know that X has the property Y but I don’t know what Y is,” or more concretely, one can represent, ”I know that this person has a name but I don’t know what it is.” The commentators then suggest that this would violate the proposed hierarchy because explicit representation of attitude (I know) is possible without explicit representation of what Y is. A violation of the proposed hierarchy would be threatened only, if explicit representation of attitude, ”I know that this person has a name,” constitutes knowledge of the person’s name without making the name and that the person has this name explicit—just like in the naming game ”cat” constitutes knowledge of the fact that the animal in front of me is a cat, without making this predication explicit. However, the proposed example simply does not constitute knowledge of this kind.

A more plausible case is the ”feeling of knowing” or ”tip of the tongue” phenomenon: ”I know this person’s name, but it won’t come off my tongue right now.” Now this complicates the picture since this phenomenon introduces a distinction between what one knows *long term* and what one knows as *instantly* available. If we construe the ”knowing”

long term then there is no threat to our hierarchy, since somewhere in the mind there is an explicit representation, ”This person's name is Susan.” If construed in terms of immediate availability, then the explicit representation ”I know that person’s name” does not constitute immediate knowledge of that person’s name, in a similar way as the representation, ”He knows that person’s name” does not constitute knowledge of that person’s name. Hence, there is no violation of the proposed hierarchy.

Nichols and Uller mount a different attack on the proposed hierarchy by showing that animals which presumably lack the capacity for explicit factuality nevertheless represent explicitly their mental state of perceiving an event, as shown in the experiments by Cowey & Stoerig (1995) on monkeys with unilateral lesions of the visual cortex. There are two ways in which our analysis can be applied to these findings.

(1) We can go along with a rich interpretation that monkeys represent events as being visual (explicit attitude) but deny the presumption that monkeys are incapable of explicitly representing factuality. What is the evidence that they can’t? One kind of evidence for such an incapacity would be the lack of pretend play. Even anecdotal evidence for such an ability in apes is scarce (Byrne, 1995) not to mention any reliable experimental evidence. However, this does not mean that primates are incapable of representing factuality. Like children with autism (Lewis & Boucher, 1988), who one does not want to deny the capacity to represent factuality altogether, they may see no point in pretence.

(2) We can accept that primates are not able to represent factuality but deny that the study by Cowey and Stoerig establishes that monkeys represent their attitude towards visual events. As Nichols and Uller’s careful formulation of ”lights” and ”non-lights” already suggests, it could be that in the second part of the experiment monkeys do not press the ”light” button because they represent that they have <u>seen</u> something, but because of the presence of some event with a certain property, e.g., something shiny (which we call light). In their blind

field they perceive lights not as shiny (hence they do not press the ”light” button) but they do perceive other properties like its position and as a button to be pressed. Or the dorsal system deals with predication implicit representations, and thus has not predicated all the distinguished features of an object or event; i.e. without the ventral system, the monkey does not perceive objects or events as coherent entities.

None of the solutions commits us to assuming that the visual system is drastically dissimilar between humans and monkeys, except for the differences that are inherent in the assumptions. In particular, if we assume (as **Nichols & Uller** seem to) that monkeys are not capable of explicit factuality then their ventral path evidently does not serve this purpose in contrast to humans. Moreover, if explicit factuality is required for consciousness then monkeys' ventral path must differ from humans in that it does not provide a conscious experience of the perceived events. In other respects, however, the functions of ventral and dorsal pathways may be the same in these species.

Nichols and Uller present a second counterargument along similar lines pertaining to declarative memory in humans and monkeys. In animals, the hippocampus seems to be responsible for creating memory of conjunctions of features that can be dealt with flexibly (Squire, 1992), but there is no evidence that it creates memories that declare something to be so. On the other hand, in people the memories formed by the hippocampus are genuinely declarative. A memory system built for dealing with one-off conjunctions in a flexible way was perhaps the most suitable starting place for evolution to mould a genuinely declarative memory system in Homo Sapiens.

Mercado & Murray, too, wonder to what extent dolphins have propositional knowledge; and, even if dolphins do not, whether they are able to represent their attitude of uncertainty explicitly. Mercado & Murray point out that dolphins choose to escape from

conditions of uncertainty. Does this indicate representation of a propositional attitude? Maybe not: The dolphin may escape uncertainty not because it has represented itself as being uncertain (i.e. not because it has attitude-explicit knowledge), but because of other effects the uncertainty has on the dolphin; e.g. aversive physiological effects. A better way of getting at attitude explicitness in animals may be to train them to respond with different levers when events happen with different long-run probabilities. Then see if the animal can transfer those responses to assessing singular events; the responses then form confidence ratings for the event happening. A similar methodology is used with children to check on their awareness of uncertainty in the context of implicit knowledge (**Ruffman**; **Goldin-Meadow & Alibali**).

Zelazo and Frye charge that we undermine our proposed hierarchy of explicitness by considering the possibility that explicit factuality might be sufficient for explicit representation of attitude (hence consciousness) because one can infer from something being a fact that one knows it to be a fact (by applying an ascent routine: Gordon, 1995). This ability to infer, however, is quite a different matter than the hierarchy of explicitness. We are not claiming that one couldn't explicitly represent "Fb is a fact" without leaving "I know...." implicit. Since we are not claiming this we do not undermine the hierarchy. Our claim is only that although there is the possibility of representing factuality explicitly and leaving the attitude of knowing implicit it may be difficult to detect actual instances of it, since people when probed can infer and then explicitly represent their attitude of knowledge for fact-explicit knowledge.

In order to incorporate consciousness into our picture we rely on the higher-order thought (HOT) theory of consciousness. Our commitment to this theory is based on the observation that we find ourselves incapable of thinking of an instance when we could say that we are conscious of some fact without the ability (requiring a higher order mental state) to specify the first-order mental state by which we behold this fact. **Zelazo and Frye** even find

this observation 'almost tautological'. Yet, when we capture the generality of this observation in the principle that it is a necessary condition for consciousness of a fact X to have a second order thought about the first order mental state with the content X, they object to it as not very compelling.

O'Brien and Opie are right, HOT is a controversial theory in the philosophy of consciousness. To a large degree this controversy is a result of relying exclusively on phenomenal intuition. This is undoubtedly an excellent starting point, but at some point of refinement introspective intuitions become inconclusive since people tend to have different intuitions (our reaction to Block's, 1995, examples demonstrating the separability of access and phenomenal consciousness). Further progress requires a theory that can make predictions for a field that goes beyond direct intuitions. Implicit and explicit knowledge is such a field, because it ties consciousness (as a particularly strong form of explicitness) in with other distinctions like directness of test, voluntary control, and hypothetical reasoning. A main purpose of our contribution is to explain how these different aspects relate to each other and form clusters—a point particularly appreciated by **Evans and Over**. HOT does not provide us with these connections, HOT only ties consciousness to explicitness of attitude and does not bear the real burden in our project as **O'Brien and Opie** claim. In fact, our project may help vindicate HOT if with its help we can make the correct empirical predictions.

Even without a HOT theory of consciousness we can bring order to the empirical facts, including those of artificial grammar learning. If one wishes to treat guesses and forced choice responses as evidence of pre-existing conscious explicit knowledge, knowledge that doesn't require HOTs (even though conscious) that's fine, its almost just a terminological issue (as **Zelazo and Frye**'s Levels of Consciousness suggest). But the facts are that in artificial grammar learning and other paradigms, subjects acquire knowledge about which they lack HOTs (is attitude implicit), and this is exactly what our framework makes clear. It also seems

quite natural to call such knowledge unconscious, even though it does affect conscious experience: Task demands lead it to affect conscious experience down stream of processing e.g. as preferences, but not as experienced knowledge.

With respect to preferences, **Bornstein** asks whether our framework only deals with propositional knowledge rather than implicit affects or motivational states. Experience can causally influence our affective and motivational states, leading to knowledge of the states' existence (I know that I have the property of being in state X, where X is e.g. liking an object), without there being knowledge of having experienced the object before (i.e. there is no retrieval of the representation formed during the perception of the object : "I see that this object has this structure"). Only a representation of the structure need be formed, not predicating it to a particular object at a particular time. For affective states to be altered by e.g. visual experience, there is no need to represent having seen the object, or to represent that the affective experience is linked to having seen the object. Thus, implicit representations can (by task demands) ultimately lead to some sort of conscious experience, that has its own attitude explicit representation associated with it even though the conscious experience is not one of knowing.

Tzelgov, Ganor, & Yehene suggest that even predication-implicit knowledge can be conscious. In support of their claim they cite the results of Tzelgov, Porat, and Henik (1997), who presented subjects with one of eight words for different durations. Subjects showed a Stroop effect only when they could report which of the eight words it was at above chance levels. This result is entirely consistent with those of Cheesman & Merikle (1984) that we discussed in the target article - the word report task is a test of objective threshold and can be performed with a predication implicit representation. Tzelgov et al claim the subject is conscious of the word because the subject reported it The subject IS conscious of the word (or is at least led to be conscious of it by the task demands) and in so doing forms a relevant

HOT (”I guess the word could have been blue”), but this occurs as an act of inference sometime after the moment of perception, so the perception itself is unconscious (no HOT involving the attitude of seeing is formed). This process corresponds to Dulany’s (1991, 1997) evocative mode of consciousness: The person becomes conscious of something but not of perception per se. Furthermore, the conscious experience arises due to the inferences caused by task demands, not directly due to the predication implicit representation formed by perception. The experiment by D.G reported by Tzelgov, Ganor, & Yehene (using the presence of semantically similar false alarms on a recognition test to indicate the lack of predication during perception) plausibly says more about memory than perception. To see this, consider their last thought experiment in which they argue a person forms a fully explicit representation. Nonetheless, if after a delay, synonyms were given as false alarms more often than control words, would Tzelgov, Ganor and Yehene argue that perception was actually predication implicit after all?

2. Explicitation

Several commentators highlighted explicitation as an important topic to address. **Georgieff and Rossetti** ask whether all implicit knowledge can be made explicit, and if not why not. Knowledge in the dorsal stream apparently cannot be. How does this figure in our scheme? As mentioned by Georgieff and Rossetti, the dorsal and ventral systems differ in more than just the implicit/explicit status of the knowledge. It is quite possible when there are independent systems like this that implicit knowledge in one system has no means of being explicated. Perhaps a crucial feature is time, as recognised by **Carlson** as well as **O’Brien and Opie**; if a representation is used by e.g. the visual system for a short period of time, it may be long enough to exert influence on subsequent processing, but too short to allow predication, factuality, etc. to be represented. Other representations whose property structure

cannot be explicated are those that carry NCC with constituent structure (thereby providing a counterargument to **Homer and Ramsay**'s claim that knowledge which is property structure implicit can be explicated by conscious reflection).

Weights in a connectionist network provide an example of this. Further, within the standard processing assumptions of connectionist networks, there is no easy means by which the representational content of weights could be represented as factual or not (Dienes & Perner, 1996). Apparent cases of procedural knowledge embedded in weights becoming explicated by reflexive abstraction (**Homer and Ramsay**) may be simply additional representations formed by hypothesis testing, rather than directly explicating the content of the implicit representations ("directly" in the sense that the mechanism predicating, etc, the property embedded in the weights is so reliable in detecting the right property it does not need to *test* whether the property is right). On the other hand, **Carlson**'s commentary contains an informative analysis of four different ways in which other implicit knowledge may be explicated.

3. Voluntary control

A few commentators noticed that involuntary processes are often associated with conscious experience. **Kinoshita**, for example, suggests that involuntary recollection poses a problem for our framework because the recollective experience implies the memory is fully attitude explicit, but, according to us, volition is also associated with full attitude explicitness. The answer is that according to our framework, a fully explicit representation is necessary for volitional retrieval but such a representation does not necessitate volitional retrieval. Thus, retrieval volition requires consciousness but conscious awareness of an item having been on the list does not require retrieval volition. Similarly, **Tzelgov, Ganor, and Yehene** point out that in a Stroop experiment, reading the word for meaning is involuntary, but the meaning still

becomes conscious. Again, on our scheme there is no reason why involuntary processes should not be conscious. We do claim automatic processes *can* be unconscious - as shown e.g. by the Stroop effects demonstrated by Cheesman and Merikle (1984).

Bibby and Underwood argue the converse point that people can invoke volitional control when their knowledge is completely implicit. Like us, they argue that people could use ”compound properties”; Bibby and Underwood describe a certain higher order property that could be used to differentially apply different grammars. The knowledge can not be completely implicit for control to happen, however; the subject has to choose one of the two grammars in some way. One way of doing this is to remember a few sequences from one of the grammars and thus use the remembered sequences to activate the right knowledge (Dienes & Perner, 1996). Thus, there would need to be explicit memory of specific items, even if there was implicit knowledge of the grammar rules. Now, let us assume that the subject has, by task demands or imagination, been able to differentially activate implicit knowledge of the two grammars. We argued that measures of familiarity (e.g. RTs to classify) could be used to indicate whether people were using implicit knowledge or strict volitional control (based on fully explicit knowledge). Bibby and Underwood show that with a two grammar design RT need not predict classification performance when implicit knowledge is being used. We agree. One doesn’t need to use second order effects (it’s unlikely that subjects use them) to make this point; e.g. subjects could think of a few g1 items to activate g1 knowledge, check the test item for g1; do the same for g2. Since the subject would do a g1 check and a g2 check each time, total RT will be the same for g1 and g2 items. On the other hand, if the subject just does a g1 check each time, RT will correlate with decisions. If the knowledge is not explicit, RT should predict classification in a one grammar design, so this provides a way of experimentally testing the explicitness of subjects’ knowledge (Buchner 1994). **Brinck** also indicates how

NCC can be applied volitionally; i.e. the knowledge can be property-structure implicit, but attitude explicit, and hence under volitional control.

Vokey and Higham suggest that implicit/explicit should be defined in terms of control rather than dissociations. We agree that control has an intimate relation with implicit/explicit, and our paper shows why there is such a relationship. We just point out that the opposition logic based on control (e.g. Jacoby, 1991) is not independent of or an alternative to ”dissociation logic”. For Jacoby, his opposition logic can only be vindicated by dissociations; it is only by obtaining clean process dissociations that one can have confidence that the equations are the right ones for isolating different processes.

Georgieff & Rossetti describe the interesting pathologies that occur when the self is not represented as agent of the action. The person suffering from e.g. schizophrenia represents themselves as observing the action (hence they are conscious of it). But they don't represent the SAS resolutions as due to the self, and hence they experience it as nonvolitional, an interesting dissociation between volition and consciousness (caused by different representations of agent of action and perceiver of action) we had not anticipated.

4. Visual perception

Bridgeman describes a recent experiment showing that information indexing particular objects can be effectively communicated from the ventral to the dorsal visual systems. These results plausibly indicate that the sensorimotor system uses predication explicit (factuality implicit) representations because particular individuals were referenced (and the information communicated between different systems). This could correspond to Bridgeman's own explanation given in his closing sentence. Alternatively, however, the ventral system may specify a region of space (not an individual) that the dorsal system can focus on (thus illustrating how different systems can communicate when one of the systems

deals with only predication implicit representations, contrary to **Goldin-Meadow and Alibali**'s strict application of our claim that predication explicitness facilitates communication between different systems). This is in fact the explanation given in Bridgeman and Huemer (1998).

5. Memory

Goshen-Gottstein argues that our theory has trouble accounting for the pattern of neuropsychological dissociations, experimental dissociations, and stochastic independence observed between direct and indirect tests of memory. In response, we make the point that one has to be careful in specifying the information required and actually accessed for a particular task. Goshen-Gottstein's first query is how we deal with a patient reported by Gabrieli, Fleischman, Keane, Reminger, and Morrell (1995), who has a lesion in the right occipital lobe leaving performance on direct tests intact, but performance on indirect tests impaired. Goshen-Gottstein points out that if the fully explicit representation is intact (supporting direct task performance), then all lower levels of explicitness have been represented, according to us, so there should be sufficient representations to support indirect task performance, contrary to the data. But in Gabrieli et al's patient, the lesion impaired only visual priming; for example, conceptual priming and auditory priming were intact. This indicates that the use of representations of visual information had been impaired, but not representations of the fact that certain words had been seen on a list. Thus, the indirect and direct tests relied on different facts (involving the visual make-up of a word versus the word itself, respectively), presenting no problem for our account.

Second, **Goshen-Gottstein** wonders how the independence of test performance on indirect and direct tests (e.g., Tulving et al., 1982) can fit in with our theory; he argues that on our account p(indirect/direct) should be greater than p(indirect). However, the fact that

p(indirect/direct) = p(indirect) requires just as much explanation even if one subscribes to independent memory systems, i.e., to explain why explicitly stored information is not accessed by indirect procedures. Thus, we can go along and admit two physiologically separate systems (also like ventral and dorsal visual paths) or have it all in one store. The explanation must lie in independence of access for indirect and direct tests, be it to separate stores or different encodings (e.g., of different information, or the same information with or without fact marker). Anderson, Bothell, Lebiere and Matessa (1998) showed how the ACT-R system, using a single interconnected set of memory chunks, can produce as much priming on indirect tests for recognised as unrecognised words. This is because the chunk storing the fact that a word was on the list can be accessed independently of the chunk relating a word and its spelling.

Finally, **Goshen-Gottstein** argues that the propositional nature of representations implies insensitivity to surface characteristics in implicit memory, but implicit memory is highly sensitive to surface features. This is a misunderstanding of our position; there is no reason why the properties represented about a stimulus and accessed by indirect tests should be restricted to the meaning of the stimulus.

Mulligan makes the related mistake of construing the difference between predication implicit and explicit simply as a matter of richness of encoding. Therefore, he argues, our analysis cannot explain why elaboration during encoding affects conceptual but not perceptual priming. However, predication implicit/explicit is not a matter of richness of encoding. Richness of encoding is a matter of property structure (rich, articulated). Thus, one could explain the dissociation between conceptual and perceptual priming because the greater elaboration produces more conceptual primes, and hence more activated material in the right part of the semantic network. Also contrary to Mulligan's claim, we do not presume that the core deficit in amnesia is a problem with conceptual-elaborative processing.

Mulligan wonders if we will be able to experimentally establish four separate states of awareness associated with memory retrieval, given that the simple distinction between Remember and Know may be reducible to two-criterion STD (Donaldson, 1996; Hirshman & Masters, 1997). However, neither of the papers cited by Mulligan as undermining the R-K distinction actually argue that the distinction should be dropped. On the contrary, both papers endorse the reality of the distinction, they just call into question *some* of the evidence for it, while approving of other evidence. We can all introspectively vindicate the difference between recollective experience and familiarity, and between volitional and nonvolitional retrieval. If you subscribe to a representational theory of mind, those different experiences must be accompanied by different representations.

6. Development

6.1. Infancy.

Poulin-Dubois and Rakosin suggest that cognitive development in infancy would be the perfect stomping ground for our theory. So, let us briefly (and very speculatively) stomp that ground to show how our analysis can be profitably applied to this field.

In the classic experiments by Baillargeon (e.g., 1987) children as infants of 4 months are not just sensitive to visual appearance but to deeper properties at the level of physical causality that Spelke (e.g., Spelke, Phillips & Woodward, 1995) has described as solidity, connectedness, spatio-temporal continuity, etc. By 4½ months children also use these "Spelke properties" to individuate objects (Spelke and Kestenbaum, 1986). When habituated to something moving behind a left screen and then without anything appearing in the spatial gap beween screens something appears from behind the right screen then infants apparently concluded that two objects must have been involved. They dishabituated more strongly to just one object being shown at test than two. This result cannot be explained by mere feature

placing of Spelke properties. The infants must have individuated different numbers of objects. However, they need not have explicitly predicated the Spelke properties to the identified individuals. The Spelke properties need only have been used to generate the appropriate number of individuals.

That infants may not explicitly predicate perceived properties to identified objects is suggested by a recent finding (Xu and Carey, 1996) involving two clearly different types of objects: a blue rubber elephant and a red toy truck. However, the two objects move alternately out from behind a screen so that on the basis of visuo-spatial continuity there could be just one object. Indeed, not before 12 months of age do infants conclude that two objects must be involved. A possible explanation is that although the younger infants use the Spelke property of continuous spatio-temporal movement to individuate a single object the additional properties of being red and a truck and at other times of being blue and an elephant are not predicated to that object. Hence they cannot derive a contradiction which would lead them to realise that two objects must be involved.

6.2. Theory of Mind and Executive Control.

Sabbagh and Clegg query the interpretation of the finding by Clements & Perner (1994) in terms of implicit knowledge of false belief. In this study children listened to a story about a protagonist who mistakenly thinks that a desired object is at location A when in fact it has been transferred to B. At about three years most children look to A in expectation of the protagonist reappearing there, but when asked where the protagonist will reappear they point to B. This has been interpreted by Clements and Perner as showing that an implicit understanding of where the protagonist will reappear (looking in anticipation) precedes an explicit understanding (answer to question). Sabbagh and Clegg suggest—in analogy to children's difficulty with deceptive responses (Carlson, Moses, & Hix, 1998; Russell, et al.,

1991)—that the young children lack the executive control to inhibit the initial predisposition to provide the usual (i.e., true) information for canonical declarative actions. Now it is not quite clear why there should be a disposition for pointing to the wrong location where the protagonist won't go (B) when asked where the protagonist will look for the object once children understand that he will reappear at A. It is also unclear why looking should be a less canonical response mode for expecting someone's reappearance (looking is not elicited as an answer to the question) than a pointing response is for answering a question.

There are however, other similar possibilities, e.g., that the pointing to B (or verbally indicating B) is not an answer to the question at all but a helpful gesture to direct the protagonist to the changed location (it is unlikely that looking would serve that purpose) and children lack the executive control to suppress these helpful pointing tendencies. This possibility is also supported by a rapidly growing literature (review by Perner & Lang, in press) showing that children's ability to answer the false belief test correctly is specifically linked to advances in executive control. For several reasons this is, however, an unlikely explanation for the results by Clements and Perner. In a follow up study (Clements & Perner, 1997) several new conditions were used (**Ruffman** mentions further controls for this finding and **Sabbagh and Clegg's** worry that everybody is transfixed on an implicit-explicit explanation is unwarranted). For instance, children had to move a mat to where the protagonist would reappear. Children who responded spontaneously moved it more often correctly to A (as often as they looked to A) than those who deliberated and moved it hesitatingly. Why should executive control fail for deliberate responses but succeed for rash responses? This result is, though, compatible with the literature on dissociations between implicit and explicit knowledge.

Another reason why the executive control explanation is not convincing are data by Hughes (1998) and ongoing research by Perner and Lang, that the strong correlation between

the standard false belief task and executive control tasks is also observed for the ‘explanation’ variant of the false belief task (Bartsch & Wellman 1989). Children observe the protagonist looking in the wrong place and have to explain why he did so. It is unlikely that children have a natural, difficult to suppress tendency to give no or wrong explanations.

An interesting question then remains why understanding false belief develops in step with improvements of executive control. Perner and Lang (in press) identified several theories in the literature to explain this fact. One of them (Perner, 1998) relates to our analysis of the implicit-explicit distinction. Although we did not develop this theory in the target article, **Zelazo and Frye** reconstruct it incompletely from the relevant but patchy parts in Section 3.4. As they point out correctly the theory makes use of two levels of control: contention scheduling (automatic control) and the supervisory attentional system (SAS). Norman and Shallice (1986) specify this distinction mainly in terms of a list of "SAS" tasks (novel actions, inhibition of existing habits, etc.) for which the SAS is required without specifying the particular information processing characteristics of the SAS. Perner (1998) drew on the distinction between action schemata as representational vehicles and their representational content and suggested that automatic control operates solely at the level of the vehicle, while the SAS directs control on the basis of representational content. Moreover, in order to represent these content specifications without creating confusion the SAS needs to mark them as something ‘desired’, which requires some minimal theory of mind. As **Zelazo and Frye** point out correctly this level should be achieved at 18 months (or even earlier—a period for which there are not enough data available).

What has not been mentioned in the target article (only in Perner, 1998) is that for certain SAS tasks, those that require ‘executive inhibition’ (Perner, Stummer, & Lang, in press) a higher level of theory of mind is needed. The SAS has to also be concerned with the fact that the representational contents are carried by causally efficacious representational

vehicles (i.e., the SAS needs to metarepresent the existence of action schemata as representational vehicles) in order to understand the need of inhibition: Prepotent action tendencies need to be actively inhibited because they make (causal efficacy) one act even though one doesn't want to act that way. The same understanding is required for the false belief task: a belief can make (causal efficacy) people look in places where they don't really want to look. For that reason—according to theory—the false belief task is mastered at the same age as executive inhibition tasks like the DCCS (dimensional change card sort) task as the data by the commentators themselves show (Frye, et al., 1995).

6.3 Gestures.

Goldin-Meadow and Alibali point out that gestural expressions of reasoning processes when, e.g., solving mathematical equations are not at the same level of implicitness as anticipatory eye movements in the false belief task (mentioned above). We agree but, perhaps, for slightly different reasons. We agree that visual orienting responses are not to be set on a par with manual gestures and that manual gestures can be put to quite different uses. In the false belief experiment the manual gesture of pointing serves as a declarative act, just like saying "there," in order to communicate the relevant information. In contrast, when solving maths problems the gestures are not intended to express or communicate anything. So we agree with the expectation for one of the proposed experiments: when making speakers aware of their gestures they will restrict them to expressing fact-explicit knowledge.

However, the predictions for the other experiment with the eye tracker seem less clear. The eye gaze measured in the false belief experiment is an orienting response: the child looks to the location (A) where the protagonist mistakenly thinks the object is because the child expects the story protagonist to make an appearance there. The looking is an integral step of how the child interacts with the story events. Without looking to A the child won't see the

protagonist. In contrast, gestures accompanying maths problems are not integral in the same sense. As **Goldin-Meadow and Alibali** point out they are ”symbolic,” certainly in the sense that they map a thought process without being part of that process. The problem with the commentators’ prediction about eye gaze patterns is that in the context of the maths problems eye gaze might serve the same purpose as manual gestures of mapping thought processes without being an integral part of them. Hence, eye gaze patterns in this task may be as much predication-explicit as manual gestures according to the commentators’ argument.

Part of their argument why the knowledge underlying gestures is predication-explicit rests on the observation that the knowledge expressed in gestures is often also used for the later generated solutions. This can be explained by a piece of knowledge getting hold of different response modalities. This kind of ”accessibility”, however, does not require predication-explicitness. Explicit predication is required when one part of the system is looking for a particular kind of information that exists in another functionally unrelated part. Hence, the more relevant evidence for predication explicitness is that children rate their solutions conforming to their gesturally expressed knowledge as more reliable than their solutions conforming to unexpressed procedures. However, these confidence ratings suggest that the gesturally expressed knowledge is not only predication-explicit but also factuality-explicit, i.e., there is some awareness of the gesturally expressed being reliable to some degree. This underlines the contrast to anticipatory looking in the false belief task according to the data presented by **Ruffman:** Children rate the solution expressed by their anticipatory looking as having zero probability (Ruffman, et al, 1998).

Alibali & Koedinger wonder in this context, what advantage accrues from thinking about procedural knowledge as a ”fact”. This question smacks of a misreading of what we mean by these terms. We are in no way concerned about whether the existence of some procedural knowledge is or is not a fact. Rather we are concerned about whether knowledge (embedded

in procedures or otherwise) represents (see point 1.5 of this Reply for refinement) a fact as a fact or not. For instance there is the fact that for y=2x+1 and x=5, y equals 11. A calculator knows this purely procedurally. Given the equation and a value for x it will spit out ”11”. It does not know that this underlying regularity is a fact, i.e., it could not pretend that the answer is 12, or provide a confidence rating for the answer ”11”. That is one reason why the ability of adults to provide higher confidence ratings for solutions conforming to gestured procedures than for solutions conforming to ungestured procedures indicates factuality-explicit knowledge.

Moreover, **Alibali & Koedinger** suggest that the findings on gestured versus verbalised knowledge of procedures can be satisfactorily modelled within ACT-R by the differential activation of declarative chunks. Weakly activated chunks may fail to fire complex language productions but may fire simpler, more well-practised productions for gestures. This suggestion strikes us as odd for the following reasons:

1. Does the fact that new and better knowledge is often expressed in gestures only, mean that children get a lot more practice on gesturing algebraic procedures than talking about them?

2. How does this modelling square with **Goldin-Meadow and Alibali's** suggestion that knowledge expressed in gestures is factuality-implicit? Is the suggestion that strength of activation represents factuality: being above a certain level of activation represents that it is a fact? If that is so, then how does ACT-R implement a well practised procedure that is deemed unreliable, or an overlearned procedure that once was explicit and then, through automatisation, has become implicit and unverbalisable again? This latter problem is a well known problem for threshold theories of consciousness (Baars & McGovern, 1996, p. 76).

7. Models of Learning

Jimenez and Cleeremans imply that knowledge need not be representational: Tuning relevant neural pathways does not form a representation, it's just a process; the corresponding claim in a connectionist network is that weights don't represent, only activation patterns do. But on our functional definition of representation, weights are representational, since they have the function to covary with various structures in the world (remember that RTK does not imply a Language of Thought). In a Hebbian network, a weight linking two nodes (representing, say the presence of A and B, respectively) has the function to indicate the covariation between A and B. That is, the weight represents that covariation. Correspondingly, the weight has all the features of a representation (Perner, 1991): It is singular (it is about A and B and not C nor D); it can misrepresent (for example, if the nodes themselves misrepresented A or B by being triggered by a C or a D on a dark night, then the Hebbian rule would lead the weights to likewise misrepresent), and so on. The weight has this content but it is nonconceptual content: It is not composed of constituents that satisfy the generality constraint, nor does it satisfy the generality constraint itself (and *typically* activation patterns carry NCC as well). Nonetheless, weights and neural pathways are representational (on a teleological account and teleological views of representation are perhaps the philosophically dominant ones). As long as Jimenez and Cleeremans accept that neural pathways have certain functions (of indicating certain contents), our framework remains applicable to the priming cases they mention.

Vokey and Higham consider other learning mechanisms - for example, the storage of instances - for which they question our use of the term implicit. Instead, such mechanisms produce knowledge that might be better described as latent knowledge, distributed over the database. We agree that the knowledge latent in exemplars is not ipso facto implicit in our sense. But we think it is the way in which the knowledge is implicit, rather than simply latent,

that captures an important part of the attraction of paradigms like artificial grammar learning[2]. The exemplars or episodes may be implicit (not predicated to a particular spatiotemporal learning context) or explicit (capable of providing recollective experience). Also, inferences based on the exemplars (in producing classification decisions, for example) may also be explicit or implicit; i.e. represented as knowledge because their appropriate causal origin is represented, or considered as mere guesses. Implicit learning, by most people's intuitions, would be said to occur when either the exemplars themselves or the inferences based on them are implicit (in our sense), not simply latent (in Vokey and Higham's sense). If the knowledge was simply latent, it leaves open the possibility that people could describe which training items they brought to mind (recollective experience) and how they assessed their similarity with the test item (justified knowledge of their grammaticality judgements).

Marescaux and Chambres indicate the complexity of the artificial grammar learning task and the range of learning mechanisms (connectionist, instance storage, etc) that may be responsible for performance. They correctly point out (as we did in the article) that confidence judgements about grammaticality judgements do not provide direct evidence about the implicit/explicit status of knowledge of the grammar per se. To do the latter, we need to infer how the knowledge is represented. Agreed, this makes life difficult but exactly the same problem exists in inferring the implicit/explicit nature of the knowledge whether one subscribes to our framework or not. If one can plausibly infer what the "rules" are (instances, n-grams, etc), then our framework enables one to test in what ways the knowledge is implicit or explicit.

Various commentators recommend the use of the ACT-R (Anderson & Lebiere, 1998) framework for understanding implicit knowledge. The accomplishments of ACT-R are indeed impressive. There is no apparent inconsistency between the ACT-R model and our

[2] Being latent captures another separate part of the attraction!

framework, and points of concordance are noted by **Lebiere and Wallach**. However, whether ACT-R is able to incorporate fully the distinctions made by our framework necessary for understanding human cognition remains to be seen. For instance, the account of explicit recognition memory (Anderson, et al., 1998; see Memory above) makes use of a context chunk that represents particular words, e.g., ”hare” as member of the learned word list. Lebiere and Wallach consider this an instance of explicit predication but the model leaves open whether the context chunk represents the complex property, ”list with ‘hare’ in it”, or the predicating proposition, ”The list has ‘hare’ in it”. Apart from the model users’ intentions, what makes the context chunk a representation of the latter rather than the former? How would the model distinguish these two psychologically different cases?

Alibali and Koedinger even suggest that ACT-R leads to predictions at odds with our theory: Our theory can not easily interpret a person being able to state a theorem but unable to apply the rule in context. This is not difficult for us - on any account, to apply the theorem the person must (a) realise its relevance; and (b) must have other supporting knowledge relevant to the problem set. The person may be lacking in (a) or (b), even if knowledge of the theorem is quite explicit. Alibali and Koedinger further wonder how our theory accounts for different types of implicitness observed in people. In their examples of different degrees of implicitness, there is a need to distinguish generality of the rule induced from its explicitness (i.e. a more general rule does not ipso facto mean more explicit), a distinction often missed in the literature. For example, in their second paragraph, the number you add to the first number to get the second increases by one in each successive number pair. This rule would be difficult to apply to pairs much smaller or larger than the pairs trained on, but there need not be anything more implicit about the rule (in our sense) than the rule $y=2x+1$.

Noelle argues that the distinction between implicit and explicit learning may be best understood in terms of different brain systems rather than different propositional attitudes. We agree that the sources of dissociations within a knowledge domain are unlikely to be purely due to content differences. In many cases, different brain regions are likely involved. However, the different brain regions can compute different contents which gives them their implicit or explicit function. Noelle doesn't confront the question of why we call the knowledge in the different systems implicit or explicit; this is where our framework clarifies. Thus, we can explain why completely different brain regions show similar dissociations, e.g., vision (parietal and temporal cortex) vs. theory of mind (prefrontal cortex).

Finally, **Gorman** considers the special case of learning involved in scientific discovery. He says he has argued that Bell followed an "implicit confirmation heuristic" (Gorman, 1995). We are not entirely sure what exactly was meant to be implicit. Bell of course doesn't say that he is following a confirmation heuristic in his notebooks; but he may just have regarded this as not something useful to put in his notebook. Similarly, for protocol analyses; they provide suggestive but not definitive evidence about which heuristics may be implicit, because subjects will only say in their protocol what they think the experimenter is interested in hearing. What was left out perhaps could be confidently reported if the subject was directly asked about it (the normal problems with relying on free report as an exhaustive measure of explicit knowledge). Nonetheless, there is plenty of scope for interesting further work on the role of implicit knowledge in scientific discovery.

7. Conclusion

The different views the commentators confronted us with have greatly expanded and clarified our own understanding of the implications of our ideas. It is reassuring that our ideas

stand up to such insightful scrutiny, and we look forward to their further development. Our final comment is for the zen-like commentary of M. J. **O'Brien.** In response, we merely raise a finger. If O'Brien raises a finger back, we will chop it off. And in that moment he will attain the attitude of enlightenment.

References

Ahmed, A., & Ruffman, T. (1998). Why do infants make A not B Errors in a search task, yet show memory for the location of hidden objects in a nonsearch task? Developmental Psychology, 34(3) , 441-453.

Anderson, J.R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. Journal of Memory and Language, 38(4), 341-380.

Anderson, J.R., & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum.

Baars, B.J., & McGovern, K. (1996). Cognitive views of consciousness: What are the facts? How can we explain them? In M. Velmans (Ed.), The science of consciousness: Psychological, neuropsychological and clinical reviews (pp. 63-95). London and New York: Routledge.

Baillargeon, R. (1987). Object permanence in 3 1/2 and 4 1/2 - month old infants. Developmental Psychology, 23 , 655-664.

Bartsch, K., & Wellman, H.M. (1989). Young children's attribution of action to beliefs and desires. Child Development, 60 , 946-964.

Bermudez, J. L. (1995) Nonconceptual content: From perceptual experience to subpersonal computational states. Mind and Language, 10, 333-369.

Block, N. (1995). On a confusion about a function of consciousness. Behavioral and Brain Sciences, 18(2), 227-287.

Bridgeman, B., & Huemer, V. (1998). A spatially oriented decision does not induce consciousness in a motor task. Consciousness and Cognition, 7, 454-464.

Brinck, I. (1997). The indexical 'I'. Dordrecht, Boston, London: Kluwer Academic Publishers.

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

Byrne, R. (1995). The thinking ape: evolutionary origins of intelligence. Oxford: Oxford University Press.

Carlson, S.M., Moses, L.J., & Hix, H.R. (1998). The role of inhibitory processes in young children's difficulties with deception and false belief. Child Development, 69(3), 672-691.

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

Chrisley, R. L. (1996). Non-conceptual psychological explanation: Content & computation. D.Phil. thesis, University of Oxford.

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

Clements, W., & Perner, J. (1997). When actions really do speak louder than words--but only implicity: Young children's understanding of false belief in action. Unpublished Manuscript, University of Sussex.

Cowey, A., & Stoerig, P. (1995). Blindsight in monkeys. Nature, 373, 247-249.

Currie, G., & Ravenscroft, I. (in press). The development of pretence. In G. Currie & I. Ravenscroft (Eds.), Meeting of minds: thought, perception and imagination. Oxford: Oxford University Press.

Cussins, A. (1992). Content, embodiment and objectivity: the theory of cognitive trails. Mind, 101, 651-688.

Diamond, A. (1985). Development of the ability to use recall to guide action, as indicated by infants' performance on AB. Child Development, 56 , 868-883.

Dienes, Z., & Perner, J. (1996). Implicit knowledge in people and connectionist networks. In G.Underwood (Ed.), Implicit cognition (pp. 227-255). Oxford: Oxford University Press.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. Memory & Cognition, 24, 523-533.

Dulany, D.E. (1991). Conscious representation and thought systems. In R. S. Wyer & T. K. Srull (Eds.), Advances in social cognition (vol. 4) (pp. 91-120). Hillsdale, NJ: Erlbaum.

Dulany, D.E. (1997). Consciousness in the explicit (deliberative) and implicit (evocative). In J. D. Cohen & J. W. Schooler (Eds.), Scientific approaches to consciousness (pp. 179-212). Hillsdale, NJ: Erlbaum.

Ebbinghaus, H. (1885). Über das Gedächtnis. Leipzig: Duncker und Humblot.

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

Fodor, J.A. (1975). The language of thought. Cambridge, MA: Harvard University Press.

Frye, D., Zelazo, P.D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. Cognitive Development, 10 , 483-527.

Gabrieli, J.D.E., Fleischman, D.A., Keane, M.M., Reminger, S.L., & Morrell, F. (1995). Double dissociation between memory systems underlying explicit and implicit memory in the human brain. Psychological Science, 6, 76-82.

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.

Gorman, M. E. (1995). Confirmation, disconfirmation, and invention: The case of Alexander Graham Bell and the telephone. Thinking and Reasoning, 1, 31-53.

Harris, P.L. (1989). Object permanence in infancy. In A. Slater & G. Bremner (Eds.), Infant development (pp. 103-121). Hove and London: Lawrence Erlbaum Associates.

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

Hirshman, E., & Master, S. (1997). Modelling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. Memory & Cognition, 25(3), 345-351.

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. British Journal of Developmental Psychology, 16(2) , 233-253.

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

James, W. (1890). The principles of psychology. London: MacMillan and Co.

Keil, F. (1998). The most basic units of thought do more, and less, than point. Behavioral and Brain Sciences, 21, 75-76.

Keil, F.C. (1989). Concepts, kinds, and cognitive development. Cambridge, MA: A Bradford Book.

Kosslyn, S.M. (1975). Information representation in visual images. Cognitive Psychology, 7, 341-370.

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

Lewis, V., & Boucher, J. (1988). Spontaneous, instructed and elicited play in relatively able autistic children. British Journal of Developmental Psychology, 6, 325-339.

Nichols, S., & Stich, S. (1999). A cognitive theory of pretense. Unpublished manuscript, College of Charleston

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

Peacocke (1993) A Study of Concepts. MIT Press, Cambridge.

Perner, J. (1991). Understanding the representational mind. Cambridge, MA: MIT Press. A Bradford book.

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., & Lang, B. (in press). Theory of mind and executive function: is there a developmental relationship? In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), Understanding other minds: Perspectives from autism and developmental cognitive neuroscience. Oxford: Oxford University Press.

Perner, J., Stummer, S., & Lang, B. (in press). Executive Functions and Theory of Mind: Cognitive Complexity or Functional Dependence? In P. D. Zelazo, J. W. Astington, & D. R. Olson (Eds.), Developing theories of intention: Social understanding and self-control. Mahwah, NJ: Lawrence Erlbaum Associates.

Piaget, J. (1945). Play, dreams, and imitation in childhood. New York: W. W. Norton.

Prinz, W. (1990). A common coding approach to perception and action. In O. Neumann & W. Prinz (Eds.), Relationships between perception and action: Current approaches (pp. 167-201). Berlin: Springer-Verlag.

Pylyshyn, Z.W. (1973). What the mind's eye tells the mind's brain: a critique of mental imagery. Psychological Bulletin, 80, 1-24.

Quine, W.V.O. (1951). Two dogmas of empiricism. Philosophical Review, 60, 20-43.

Roberts, P.L., & MacLeod, C. (1995). Representational consequences of two modes of learning. The Quarterly Journal of Experimental Psychology, 48A(2), 296-319.

Ruffman, T., Clements, W.A., Import, A., & Connolly, D. (1998). Does eye direction indicate implicit sensitivity to false belief? Unpublished manuscript, University of Sussex

Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991). The 'windows task' as a measure of strategic deception in preschoolers and autistic subjects. British Journal of Developmental Psychology, 9, 331-349.

Spelke, E.S., & Kestenbaum, R. (1986). Les origines du concept d'objet. Psychologie Francaise, 31, 67-72.

Spelke, E.S., Phillips, A., & Woodward, A.L. (1995). Infant's knowledge of object motion and human action. In D. Sperber, D. Premack, & A. J. Premack (Eds.), Causal cognition. A multidisciplinary debate (pp. 44-78). Oxford.

Sperber, D. (1997). Intuitive and reflective beliefs. Mind & Language, 12(1), 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.

Tulving, E., Schacter, D.L., & Stark, H.A. (1982). Priming effects in word fragment completion are independent of recognition memory. Journal of Experimental Psychology: Learinign, Memory & Cognition, 8, 336-342.

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

Tzelgov, J., Porat, Z., & Henik, A. (1997). Automaticity and consciousness: Is perceiving the word necessary for reading it? American Journal of Psychology, 110, 429-448.

Xu, F., & Carey, S. (1996). Infants' metaphysics: The case of numerical identity. Cognitive Psychology, 30, 111-153.