Metarepresentation in Philosophy and Psychology

Sam Scott (sscott@ccs.carleton.ca)
Department of Cognitive Science
Carleton University, Ottawa, ON K1S 5B6

Abstract

This paper brings together two definitions of metarepresentation: Dennett's notion of metarepresentation as second-order representation, and an alternative definition of metarepresentation found in the work of Leslie, Frith, and Baron-Cohen on autistic children. I show that the two definitions are not in any way compatible with one another, and that the assumption that they *are* compatible can lead to confusion about the nature of higher cognition. I illustrate this potential for confusion through the analysis of some claims made in a paper by Whiten and Byrne on primate cognition.

Representation

I will use the term "representation" to mean *mental representation* as defined in Von Eckardt's (1999) MITECS entry. Her definition of mental representation is (I hope) sufficiently broad and uncontroversial to be acceptable to most of the various competing currents in cognitive science. According to Von Eckardt, a (mental) representation has four important aspects: "(1) it is realized by a representation bearer; (2) it has content or represents one or more objects; (3) its representation relations are somehow 'grounded'; (4) it can be interpreted by (will serve as a representation for) some interpreter." (p. 527) Points (1) and (4) in the above establish that a (mental) representation requires a subject that both bears and can interpret the representation.

Point (2) establishes what the representation can be about. The point about representing one or more objects is fairly clear, but the point about "having content" needs some unpacking. Fortunately, Von Eckardt does that unpacking for us. A (mental) representation is something that can stand for "concrete objects, sets, properties, events, and states of affairs in this world, in possible worlds, and in fictional worlds as well as abstract objects such as universals and numbers; that can represent both an object (in and of itself) and an aspect of that object (or both extension and intension); and that can represent both correctly and incorrectly." (p. 527) Von Eckardt's list is probably not exhaustive, but it does cover the ability of cognitive systems to "think about" objects in the world, counterfactual situations, and propositions and predicates, all under the umbrella term *representation*.

The only point that remains undeveloped in Von Eckardt is point (3), which states that relations must be "grounded". I take that to mean simply that there must be an external referent of some kind for any representation, although this "external" referent may only exist in a possible or fictional world.

Metarepresentation

The prefix "meta" can mean a number of different things in different contexts (e.g. "metaphysics", "metaphilosophy", "metamorphosis" to name but a few) but the usual sense attributed by philosophers is that a metarepresentation is a higher-order representation of some kind. That is, a metarepresentation is a representation of a representation. Following Dennett (1998), it stands to reason that if a representation exists as an object in the world, then it too can be represented. Dennett's examples of metarepresentation tend to be of a hybrid nature. For instance a drawing on a piece of paper is a type of non-mental representation, which is represented in the mind of the person viewing it. The mental representation is of the drawing, but since the drawing is itself a representation, the viewer has a (mental) metarepresentation of whatever it is that the drawing represents.

Despite the drawing being an "external" rather than a mental representation it does share many of the properties of the latter. Following Von Eckardt as quoted above, the drawing: (1) has a representation bearer (the paper); (2) has content (whatever the drawing represents); (3) has a referent; and (4) can be interpreted by some interpreter. Most of Dennett's examples are to do with hybrid metarepresentation – mental representations of external representations. An interesting question is whether hybrid metarepresentation is the same sort of thing as purely mental metarepresentation. Some would say no, arguing that in the hybrid case, there is a difference of content – the external and the mental represent in different ways, therefore a representation of an external representation has a different type of content from a representation of a mental representation. Dennett would not want to take this approach, opting for an intentional stance in which he could avoid discussing matters such as internal content – things represent if we can sensibly treat them as representing, regardless of whether the representation has any further degree of reality. For the

current discussion, I would like to leave aside issues of content and the differences between hybrid and purely mental metarepresentation. In any case, this paper discusses purely mental metarepresentation almost exclusively. I hope I can safely take from Dennett the intuitively satisfying notion that the definition of "metarepresentation" corresponds roughly to the definition of "higher-order representation".

In addition to being intuitive, the "higher-order" definition of metarepresentation is also the one that has seen most use in philosophical circles. Unfortunately, in a large part of the psychological literature, it is not clear that this is the definition that is in use. In what follows, I will show that the so-called "metarepresentational conjecture" that is postulated to explain certain aspects of autistic behavior is making use of a very specific and technical definition of the word "metarepresentation". Of course this fact on its own should be neither surprising nor cause for alarm – any group of scientists should always feel free to redefine terms in technical ways that suit their needs. But unfortunately, the different definitions have lead to confusion even within the psychological literature. Before moving on to this literature, however, it is worth spending some time sorting out potential confusions lurking within the definitions articulated above.

What Metarepresentation is NOT

First of all, a representation can *contain* other representations without being a metarepresentation. For instance, consider the representations that might be necessary to entertain the thought corresponding to the following proposition:

(1) Mélissa's dog is dead

At the very least, we need a representation of Mélissa's dog. We will also need a representation of some one-place predicate DEAD. Finally, it is possible that we would also need a representation of the saturated predicate DEAD(Mélissa's dog). Depending on your personal biases (i.e. connectionist or classical), you may therefore want to assert that understanding sentence (1) requires a representation of Mélissa's dog which is *contained within* a representation of the predicate DEAD(Mélissa's dog). If so, this is not the same thing as the representation of DEAD(Mélissa's dog) being (even partially) a metarepresentation *of* Mélissa's dog. That is, there is nothing necessarily metarepresentational going on in this situation.

So far so good, but the next assertion may be more controversial. Second-order beliefs and desires do not necessarily require metarepresentations either. To see why, consider the following first-order belief:

(2) Mélissa BELIEVES that her dog is dead

This first-order belief requires Mélissa to have a mental representation of the proposition "my dog is dead" and believe that the proposition is true (perhaps it is marked as "true" in her mental database, or perhaps she has it in her "belief box", or whatever). The important observation here is that Mélissa need not be aware of her belief. If she were, she would require a representation of it, but to simply hold the belief, no such special mental machinery is required. She need not think to herself "I believe my dog is dead" in order to believe her dog is dead. She just needs to believe that her dog is dead. Thus "believe" is a definitional label we apply to any state of affairs in which someone holds a proposition to be true. This is why we can speak of animals having beliefs, even if we are not comfortable with the notion that they may be aware of them.

Now consider the following second order belief:

(3) Anne BELIEVES that Mélissa BELIEVES that her dog is dead

What kinds of representations do we need to ascribe to Anne in this case? First, she needs the representation of Mélissa's dog, the predicate DEAD, and so on. What she doesn't need is a representation of *Mélissa's representation of* her dog, the predicate DEAD, and so on. That is, she doesn't need a second-order representation of any of these things. She can get by with her own first-order representations. But it would appear that Anne also needs to have a representation of Mélissa's BELIEF. That is to say, she needs a representation of Mélissa's mental state of believing in a way that Mélissa does not. She must be aware of Mélissa's BELIEF, while Mélissa need not be. If we consider Mélissa's mental state of believing to be an object in the world, then this mental state must be represented somehow in Anne's belief. The question of whether we need a metarepresentation here hinges on whether Mélissa's belief state counts as a representation. But as I pointed out above, for Mélissa to simply *have* the first order belief, no first order representation of belief is required. Since neither Mélissa nor Anne has any particular need of belief representation in order to be a believer, Anne's representation of Mélissa's belief need not be second-order.

So what *does* Anne require in order to hold belief (3) above? It would seem that certain processing requirements are necessary to be able to form such complex thoughts. (Recall that what she actually believes corresponds to sentence (2) above, and not to sentence (3).) First of all, Anne must be able to perform some kind of *propositional embedding*. She needs to be able to represent Mélissa's belief as a proposition with two arguments: a representation of Mélissa, and a representation of the proposition that she believes.

Furthermore, Anne needs to be capable of dealing with *referential opacity*. She must be able to remain agnostic about the truth-value of the embedded proposition ("Mélissa's dog is dead") and recognize that it has no effect on the truth-value of the belief proposition.

"Metarepresentation" in Autism Research

A particular definition of "metarepresentation" has played a very important role in research on Autism, where researchers have proposed the existence of a metarepresentational module to explain some of the deficits that autistic people exhibit. Alan Leslie, along with Simon Baron-Cohen and Uta Frith, are the principal proponents of metarepresentational modules in the psychological literature (Leslie, 1991; Baron-Cohen, 1991). Leslie in particular has put forward the *metarepresentational conjecture*: "Autistic children are impaired and/or delayed in their capacity to form and/or process metarepresentations. This impairs (/delays) their capacity to acquire a theory of mind." (Leslie, 1991, p. 73) Before dissecting what Leslie means by "metarepresentation", let's take a quick look at the evidence on which this statement is founded.

The three most classic experiments on autistic children are the picture sequencing task, the Sally/Anne task and the Smarties task, all of which reveal a selective deficit in autistic children in understanding false beliefs. For space reasons, I discuss only the Sally/Anne task (see Leslie, 1991 for the others). In this experiment dolls are used to act out a scenario in which Sally hides a marble in a basket and leaves the room. While she is gone, Anne enters and transfers the marble to a box. Sally returns, and the children are asked, "Where will Sally look for her marble?" Autistic children consistently make the incorrect prediction that Sally will look in the box. They fail to realize that in the absence of new information, Sally will retain her (now false) belief that the marble is still in the basket – to use the common term, autistic children lack an adequate Theory of Mind.

In the first act of the puppet show, the child presumably believes the following:

(4) The marble is in the basket, and
(5) Sally BELIEVES that the marble is in the basket.

Then in act 2, the child learns that:

(6) The marble is in the box

and presumably updates her beliefs incorrectly to infer that:

(7) Sally BELIEVES that the marble is in the box

Recall the conclusions of the previous discussion: 1) second-order beliefs do not necessarily require metarepresentations (it is only necessary to have the ability to *represent* first order beliefs in order to *have* second-order beliefs), and 2) propositional embedding and referential opacity are required for second-order beliefs. Following from these conclusions, it seems clear that the Sally/Anne test does not imply an autistic deficit to do with second-order representations. Rather, it implies that either: 1) the autistic child does not have a concept of belief, or 2) the autistic child has a concept of belief but cannot handle the processing requirements of referential opacity and/or propositional embedding. In fact, the evidence quoted in (Leslie, 1991) is insufficient to distinguish between these two possibilities. Children are never directly asked about the beliefs of others ("Where does Sally *think* her marble is?") Rather, they are asked something like, "Where will Sally *look* for her marble?"

The second area of evidence quoted by Leslie is the apparent lack of pretend play in autistic children, and it is on this basis that he develops the metarepresentational conjecture and defines what he means by "metarepresentation". "I have used the term 'metarepresentation' in a specific sense: to mean (e.g., in the case of understanding pretence-in-others) an internal representation of an epistemic relation (PRETEND) between a person, a real situation and an imaginary situation (represented opaquely)…" (Leslie, 1991, p. 73) This definition doesn't sound at all like the definition of metarepresentation as higher-order representation pursued above. It seems like a highly technical redefinition of the word. This is a fact that Leslie seems to be quite aware of, as he says in a footnote that "'metarepresentation' can mean something like 'a kind of proprietary (internal) representation in ToM mechanisms' and something like 'a particular concept of representation which someone grasps'." (p. 77) It is not clear what Leslie's second possibility in the above refers to, but what he probably has in mind is Perner's (1991) account, which differs from both Leslie and Dennett. Unfortunately, he does not elaborate any further. From now on, I will call the definition of "metarepresentation" as higher-order representation "metarepresentation$_1$", while Leslie's version will be "metarepresentation$_2$" (and I'll forget about Perner's definition for the purposes of this discussion).

With that in mind, let's take a look at Leslie's formalism of the PRETEND example. In his view, the predicate PRETEND (which is supposed to behave similarly to BELIEVE and DESIRE) works something like this:

(8) Mother PRETEND the empty cup "it contains tea" (p. 73)

In addition to the new definition for metarepresentation$_2$, Leslie is also using a very different formalism for his psychological predicates – three arguments instead of two. Two questions immediately arise: 1) is Leslie's formalism plausible and/or compatible with the BELIEF/DESIRE formalism pursued above? and 2) putting aside Leslie's metarepresentation$_2$, is there anything metarepresentational$_1$ in his alternative formalism?

The Plausibility of Leslie's Formalism

Much of what I have to say in this section and the next parallels critiques from Pernerwith which I am in broad agreement (for example, Perner, 1991) . Rather than give a full analysis of Leslie's ideas, I will concentrate on the points I need to make for the discussion to follow.

The first observation is that there appears to be an important difference between pretending and believing, so we need to be cautious about generalizing from one to the other. Although it is possible to have beliefs without any representation of belief, it is not at all clear that this also holds for pretence. The possibility of pretending that something is true without being aware that one is doing so seems unlikely. So whereas to believe that the cup is empty does not require the self-conscious reflection that

(9) I BELIEVE that the cup is empty,

there is no way to pretend that the cup contains tea without self-conscious reflection by the subject on her own mental state. That is, the subject would have to BELIEVE:

(10) I PRETEND that (the cup contains tea).

Therefore, unlike beliefs and desires, being able to pretend seems to imply the ability to understand pretence in oneself, and thus in others, since the forms are the same. For instance believing that:

(11) Mother PRETENDS that (the cup contains tea)

requires exactly the same representational capacities as believing that one is pretending oneself.

Getting back to the substance of the issue, Leslie's formalism is actually quite different from the above. In his system, pretence is represented more like this:

(12) Mother PRETENDS (the empty cup) ("it contains tea")

That is, he has three elements: the subject (Mother), the real situation (the empty cup), and the pretend situation (it contains tea). But why is it that in order to understand pretence, you must be aware of exactly how the real situation differs from the imagined one? In reality, you can simply say "Mother pretends that the cup contains tea" and remain unsure of whether the cup is empty, contains orange juice, or whatever. That is, you do not need to know the "real" situation to understand the pretence. All you need to know is the fact, contained in the semantics of PRETEND with it's implied referential opacity, that the real situation must differ in some way from the imaginary one. If this is clear in the case of PRETEND, it is even more so in the case of BELIEVE. It would be much too restrictive to suppose that BELIEVE requires knowledge of the actual situation as in:

(13) Mélissa BELIEVES (her dog is dead) ("her dog is dead"), or
(14) Mélissa BELIEVES (her dog is **not** dead) ("her dog is dead")

Again, the information one needs is bound up in the semantics and referential opacity of BELIEVE. When you believe that *p*, *p* may or may not be true. Further problems arise when we try to embed Leslie's formalizations of psychological predicates to form second order beliefs. For instance, to believe (12) above would require:

(15) I BELIEVE [Mother may or may not PRETEND (the empty cup) ("it contains tea")]$_a$ ["Mother PRETENDS (the empty cup) ("it contains tea")"]$_o$

where the subscript "a" above marks the actual situation and "o" marks the referentially opaque proposition. This situation just seems unnecessarily complicated. You simply don't need to know the real situation in order to evaluate the truth of psychological predicates. The principle of referential opacity gives you everything you need to know – that the embedded proposition may or may not be true regardless of the truth-value of the psychological predicate.

Metarepresentation$_1$ in Leslie's Formalism

Is there anything metarepresentational$_1$ in Leslie's formulation of the semantics of psychological predicates? Leslie has made the unusual move of including the actual situation alongside the imagined situation in his formulation of at least one of the psychological predicates. Does this move change anything in the analysis of metarepresentations$_1$ in psychological predicates? The only real complication here is the introduction of dual representations for the same object – for example, the cup as an empty cup and the cup as a cup with tea in it. This dual representation

is well accounted for in Von Eckardt's (1999) definition of mental representation. The first refers to a concrete object and/or a property of a concrete object, while the second refers to an object/property in a possible or fictional world, or in the case of BELIEVE may simply represent an object/property incorrectly. So the dual representation does not imply metarepresentation$_1$.

One other aspect of Leslie's formulation deserves consideration. In the "Mother PRETENDS" example above, the first situation (the empty cup) is referred to again in the imaginary situation (it contains tea). The anaphoric reference in the imaginary situation ("it") could perhaps be taken to imply that the *representation* of the empty cup, rather than the empty cup itself, is the subject of the imaginary representation, thus making the latter a metarepresentation$_1$, but this is probably not the interpretation Leslie had in mind.[1] In fact it is hard to imagine how such an interpretation could be made coherent, since unpacking the imaginary situation would lead to "(the representation of the empty cup) contains tea" – and that can't be right.

Metaconfusion

Leslie is self-consciously using a technical definition for metarepresentation$_2$ that does not intersect in any way with Dennett's metarepresentation$_1$. Nevertheless, for other authors, the distinction may not be so clear. The potential for confusion is quite neatly demonstrated in a paper by Whiten and Byrne (1991). In an otherwise excellent article about the implications of Leslie's ideas for studies of pretend play in primates, they explicitly state that Leslie's metarepresentation$_2$ is second-order representation (i.e. metarepresentation$_1$). But the confusion doesn't stop there. They go on to offer a summary of Leslie's theory of metarepresentation$_2$ that is worth quoting at length.

"Leslie argues convincingly that the isomorphism between the properties of mental state terms and those of pretend play is not coincidental, but signifies a fundamental psychological achievement which can generate both pretence and an ability to represent the mental states of others. What these two share is that they are *representations of representations* – labeled variously as second-order representations (Dennett) or metarepresentations (Pylyshyn, Leslie).

"In the case of mental state terms, what 'second-order' means is fairly obvious: the child's mind represents a mental state in another's mind, *believing* (for example) that her father *thinks* there is a mouse behind the chair.

"In the case of pretence, the implication is less obvious. The key point is that in pretence, as strictly defined by Leslie, two simultaneous representations of the world must coexist in a precise relationship. When a child talks into a banana as if it were a telephone …the child has a primary representation of the object as a banana and, simultaneously, a representation of it as a telephone … The pretend representation is coded or marked off in some way as metarepresentational…" (Whiten and Byrne, 1991, p. 269, their italics.)

The first two paragraphs above demonstrate the confusion nicely. The authors are explicitly running together Dennett's metarepresentation$_1$ with Leslie's metarepresentation$_2$. Furthermore, they are committing the error of assuming that second order beliefs require second order representations. To see this, consider their example in the final paragraph, which makes use of the psychological predicate THINK. They make it seem like the child must have a representation of her father's thoughts, which of course consist of representations. Therefore the child must be engaging in second-order representation, or metarepresentation$_1$. But "thinks" in this context means the same thing as "believes", and so the appropriate formulation is actually:

(16) The child BELIEVES that her father BELIEVES that there is a mouse behind the chair.

This is straight-up second-order belief, which I have shown to *not* necessarily involve second-order representation, or metarepresentation$_1$.

The final paragraph picks up on the "real situation" vs. "imaginary situation" component of Leslie's formulation and reads into it another sense of "metarepresentation", which I'll call "metarepresentation$_3$" – definition: a representation of a counterfactual state of affairs. But counterfactual representations are fully compatible with the fairly non-controversial theory of first-order mental representations put forth by Von Eckardt (1999).

The confusion in Whiten and Byrne really comes to the fore in their concluding sections, where they talk about a "cluster of metarepresentational capacities." The first capacity they discuss is indirect sensorimotor coordination – the ability that humans and some other primates have to direct the actions of parts of their bodies by looking in a mirror or at a video image of the body parts they are trying to control. This, according to Whiten and Byrne, requires "a capacity to represent the remote representation of parts of self available in the mirror or video image: second-order representation" (p. 279). This ability is straightforwardly metarepresentational$_1$ in the Dennettian sense. In fact, it is a case of hybrid metarepresentation$_1$, requiring a mental representation of an external representation.

The other two "metarepresentational" abilities are tool use and insight. Tool use (in this case, a chimpanzee using a branch to probe for termites)

[1] Recall Leslie's definition above: "…an internal representation of an epistemic relation…"

apparently requires "a capacity to generate, simultaneously with the primary perception of the branch as branch, a metarepresentation of it as probe" (p. 280). Insight is a leap from pretence to re-description as in, "'I *pretend* this rock is a hammer' … 'Aha, I could *use* this rock as a hammer'...." (p. 280, Whiten and Byrne's italics). This is much closer to Leslie's technical definition of metarepresentation$_2$ in which the representation of the world as it really is coexists with a pretend representation of the world. But as I argued above, Whiten and Byrne appear to have drawn on the occurrence of a counterfactual in Leslie's formalism to build a third sense (metarepresentation$_3$), which is at work in the above.

In conflating metarepresentation$_1$ with their own interpretation of metarepresentation$_2$ (metarepresentation$_3$), Whiten and Byrne have made two mistakes, one of which comes directly from Leslie, and one that is not explicitly present in (Leslie 1991). In the former case, they have imported Leslie's notion that psychological predicates require an explicit representation of how the world actually is in addition to the representation of how the world is believed, pretended, or desired to be, and used it to unwittingly arrive at a new definition of metarepresentation$_3$. But they have also made another mistake in equating Leslie's metarepresentation$_2$ with Dennett's metarepresentation$_1$, even to the point of citing Dennett and Leslie in the same sentence.

To be fair to Whiten and Byrne, their dissection of Leslie is itself an attempt to criticize and make some new distinctions. For instance, they point out that not all pretend play involves a real object. Humans and other apes appear quite capable of having imaginary friends, and interacting with imaginary objects. In this case, it is difficult to see what the "real situation" component of Leslie's formulation would amount to, and is evidence for at least sometimes abandoning it in favor of two-place intentional predicates. But confusion over the two original senses of metarepresentation, and the unwitting introduction of yet a third sense really manages to confuse the issue. For instance, in summing up, they speculate that perhaps, "what convinces those who interact intensively with them that chimpanzees are 'intelligent' is a facility in second-order representation." (p. 280) This is a nice parsimonious account, but it is built by equating three different definitions of metarepresentation based on a number of confusions about the nature of psychological predicates. As I have attempted to show, second-order beliefs and desires as well as the pretend play studied in Whiten and Byrne's work require propositional embedding and referential opacity, but do not necessarily require second-order representations (metarepresentation$_1$).

Conclusions and Prospects

In Dennett's *Making Tools for Thinking* (Dennett, 1998), he invites us to speculate along with him on the difference between what he terms "florid" and "pastel" representations. Florid representations are those that become explicit as objects in the world, by being encoded in language or some other physical medium (drawings on paper, for instance.) He notes that the capacity to form florid representations seems to imply the ability to manipulate the representations themselves, which leads him to raise the slogan "no florid representation without metarepresentation." He further speculates that "belief about belief" may not be the same thing at all as "thinking about thinking" – that is, having the ability to self-consciously reflect, compare notes with other thinkers, and so on. The considerations in this paper may help to shed a little light on all of these questions.

If I am right that second-order belief does not require metarepresentation$_1$, and Dennett is right that thinking about thinking requires florid representations and therefore metarepresentations$_1$, then maybe we do have the basis for a nice account of one possible difference between humans and other apes – a capacity to form and manipulate higher-order representations (that is, metarepresentations$_1$).

References

Baron-Cohen, Simon (1991). Precursors to a theory of mind. Andrew Whiten (Ed.), *Natural Theories of Mind: Evolution, Development, and Simulation of Everyday Mindreading*. (pp. 233-252). Oxford: Blackwell.

Dennett, Daniel (1998). Making tools for thinking. Dan Sperber (Ed.) (2000). *Metarepresentation*. New York: Oxford University Press.

Leslie, Alan (1991). The theory of mind impairment in autism. Andrew Whiten (Ed.) *op cit. (*pp. 63-78)

Perner, Josef (1991). *Understanding the Representational Mind*. Cambridge, Massachusetts: MIT Press.

Von Eckardt, Barbara (1999). Mental representation. Robert A. Wilson and Frank C. Keil. *The MIT Encyclopedia of the Cognitive Sciences*. (pp. 527-529). Cambridge, Massachusetts: MIT Press.

Whiten, Andrew, and Byrne, Richard W. (1991). The emergence of metarepresentation in human ontogeny and primate phylogeny. Andrew Whiten (Ed.) *op cit.* (pp. 267-282.)