In D. Sperber (Ed.), *Metarepresentation: A Multidisciplinary Perspective*. (pp. 197-223). Oxford: Oxford University Press

How to acquire a 'representational theory of mind'

Alan M. Leslie[1]

Department of Psychology and
Center for Cognitive Science
Rutgers University

The study of cognitive development is dominated by the view that concepts are essentially *packets of theory-like knowledge* (Carey, 1985, 1988; Keil, 1989). This view has emerged from a long tradition of viewing concepts as descriptions of one kind or another, though there have been and continue to be many variations and disagreements concerning the character of the associated knowledge (e.g., Murphy & Medin, 1985; for critical reviews of this tradition, see Fodor, 1998; Kripke, 1972; Laurence & Margolis, in press). The essence of this family of views is that the knowledge packet associated with the concept determines what in the world a given concept refers to or designates — it fixes what the concept is a concept of. For example, the concept DOG[2] might be associated with a knowledge structure that specifies "HAIRY, FOUR-LEGGED, ANIMAL, BARKS, WAGS TAIL, . . ." If this specification captures the structure of the concept, DOG, then this is the specification you will need to know in order to possess the concept, for the following reason: this specification, when applied to the world, is what will selectively pick out the things that are *dogs* and thus link DOG with *dogs*.

The 'concept as knowledge' view is so deeply entrenched that it is hard to see how there could be an alternative. The view has two powerful implications for conceptual development. First, the acquisition of a concept must be the acquisition of the critical knowledge that defines the concept. Second, the innateness of a concept must be the innateness of the critical knowledge that defines the concept.

Perhaps the knowledge view of concepts will prove to be correct. However, to date, there is not a single concept for which a detailed model of the critical knowledge has been worked out and empirically substantiated; there is not a single concept whose acquisition or innateness has been understood. All conclusions therefore remain highly tentative.

Much of the most interesting work in cognitive development over the last twenty years has been concerned with abstract concepts, that is, with concepts that are not reducible to sensory transduction. Many abstract concepts are now thought to emerge early in development. Mental state concepts, such as BELIEVE, DESIRE, and PRETEND, are among the most abstract we possess. It is striking that these concepts are routinely acquired by all normally developing children before they attend school and are even acquired by children who are mentally retarded. The verbal labels associated with these concepts are never explicitly taught, yet are typically in use around the third birthday; by contrast, words for colors, a salient sensory property, very often are explicitly taught by parents, but are typically not learned any earlier and are often learned later. Mental state concepts provide a crucial challenge to our attempts to understand

[1] *Acknowledgments*: I am grateful to the following friends and colleagues: Eric Margolis and Susan Carey for helpful discussions, and to Jerry Fodor, Shaun Nichols, and Brian Scholl for helpful discussions and detailed comments on an earlier draft.

[2] I use small caps when referring to a concept as opposed to what the concept denotes (italicized). Normally, one could simply say that the concept is a psychological entity, while what it denotes is not, e.g., DOG refers to *dogs*. But in the case of mental state concepts what they denote are also psychological entities.

what is required for the acquisition and possession of abstract concepts. In our attempts to understand early emergence, one variant of the knowledge view of concepts has become popular; in this variant, critical knowledge is said to take the form of a *theory*. The concept BELIEF has been a central focus of these attempts.

At first sight, it is plausible that the acquisition of the concept BELIEF must be theory formation because how else can we come to know abstract things, if not by employing theories. The so-called 'theory-theory' of BELIEF has gained a widespread credence (Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994, 1995; Perner, 1991; Wellman, 1990). However, I believe that current attempts to develop a theory-theory of BELIEF have foundered. In this chapter, I will explore the reasons for these present difficulties. Because I have elsewhere written extensively on the relevant experimental evidence and developed an alternative framework to theory-theory (see e.g., Leslie, in press for a review), here I shall confine myself to examining the deeper motivations for theory-theory in order to say why I believe the entire enterprise is mistaken.

Three versions of theory-theory

There seems to be about three versions of 'theory-theory' currently active; they are not always clearly distinguished, though they need to be. The first is simply the idea that not all knowledge is sensory in character and that some knowledge is concerned with 'understanding' the world. This seems sensible and true. To say that people acquire commonsense 'theories' in this sense is just to say that they acquire abstract knowledge and opinion. For example, people develop opinions about the existence of ghosts (Boyer, 1994), the nature of consciousness (Flavell, Green & Flavell, 1993), and the disposition of heavenly bodies (Vosniadou, 1994). People also develop opinions about circumstances which will cause beliefs to be false. This might be called a 'representational theory of mind' and, if so, I shall argue that the concept BELIEF is prior to the theory.

A second current version of theory-theory is more controversial. This view holds that routine early cognitive development and the process of scientific discovery both result in knowledge of 'theories;' in particular, it is claimed that the child's 'theory of mind' really *is* a theory. I will discuss this version in the next section where I conclude that it is not useful to insist that things which are merely theory-*like* really *are* theories.

The third version of theory-theory goes deeper than the first two because it tries to account for the nature and acquisition of concepts. In its most explicit and sophisticated form, developed by Carey (1985, 1988), fundamental structures of thought are said to depend upon 'ontological' concepts, such as PHYSICAL OBJECT, LIVING THING, and so forth. The identity of an ontological concept is determined by the role it plays in a set of explanatory principles grasped by the child. A given set of explanatory principles is domain-specific and theory-like, but, most importantly, constitutes the 'packet of knowledge' that allows the child (or other user) to pick out just those things in the world to which the concept refers. Put more simply, a concept, e.g., DOG, is possessed by grasping a certain commonsense theory, namely, the theory that tells the user what kind of thing a *dog* is. Acquiring this concept is acquiring the theory of what a *dog* is. If (knowledge of) a given theory is innate, then the associated concept will also be innate; if a given theory must be acquired, then the associated concept must be acquired (by acquiring knowledge of the theory). Perner (1991, 1995) has applied this framework to the concept, BELIEF. In his account, the child acquires the concept BELIEF by acquiring a theory of what *beliefs* are, namely, the theory that *beliefs are representations*. I discuss this version of theory-theory in a later section, pointing out that it requires the child to have obscure knowledge for which there is no independent evidence and that it still fails to account for possession of the concept, BELIEF.

Some current beliefs about BELIEF

The empirical basis of the belief problem is as follows. Wimmer and Perner (1983) developed a test of false belief understanding (the Maxi task) which showed that the majority of six-year-old children could pass, while four-year-olds performed at chance. Baron-Cohen, Leslie and Frith (1985) subsequently modified this task, simplifying it (the Sally and Ann task, Figure 1). They found that the majority of normally developing four-year-old children passed this version. This study also found that a majority of mildly retarded children with Down's syndrome could pass the task, but that children with autism, even with normal IQ's, failed. Subsequently, numerous studies have confirmed and extended these results (for reviews, see Happé, 1995 and Leslie, in press). By age four, most normally developing children are demonstrably employing the concept BELIEF.

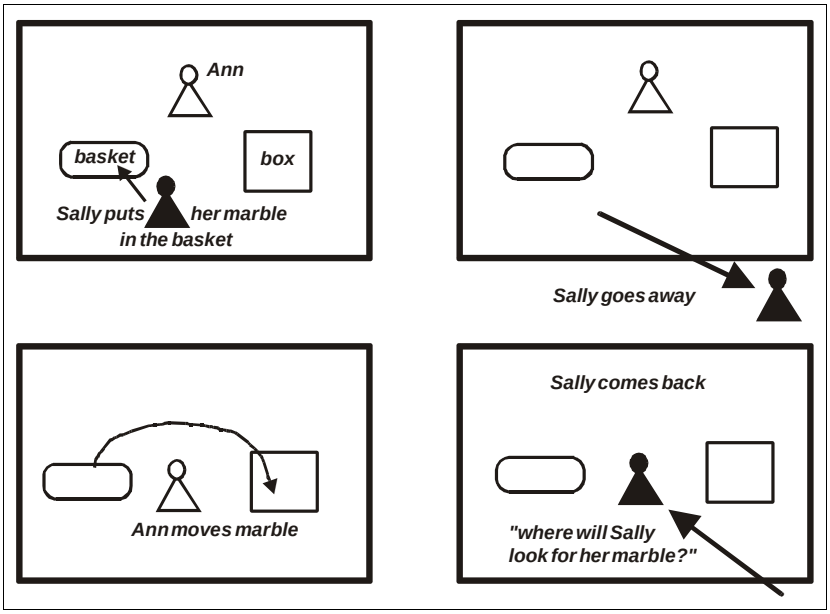

Figure 1 A standard test of false belief attribution. In addition to the prediction question shown here, children are asked two control questions, one to check that they remember where Sally put the marble and one to check they realize where the marble is currently. An alternative to the prediction question is the *think* question: Where does Sally think the marble is? Both prediction and think questions yield similar results with normally developing children and with children with a developmental disorder. (After Baron-Cohen, Leslie, & Frith, 1985).

The real-theory-theory

One version of theory-theory is that people, including children, 'have theories.' As I indicated, there is little in this claim to disagree with, in part because the notion of 'theory,' especially when extended from science to commonsense, is vague enough to cover almost any kind of knowledge and opinion.

Recently, however, the claim has been pushed to an extreme in which routine cognitive development and the process of scientific discovery are claimed to be essentially identical (e.g., Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1995).[3] Although there is some disagreement within the theory-theory camp over whether the child-as-scientist claim relates to the *process* of development (denied by Wellman, 1990 and by Perner, 1995 but espoused by Gopnik and Wellman, 1995 and by Gopnik and Meltzoff, 1997) or only to the *outcome* of that process (Perner, 1995; Wellman, 1990), there appears to be agreement that it relates at least to the outcome. Gopnik and Wellman (1994, 1995) develop their claim by thinking of scientific theories as a species of psychological entity. They are not concerned with the substance of any particular scientific theory, but rather with the general psychological properties of that whole class of knowledge. From this point of view, they generate a list of critical properties. The critical properties of scientific theories are said to be abstractness, coherence, predictiveness, defeasibility, interpretation of evidence, and explanatoriness. Gopnik and Wellman then point to features of the child's 'theory of mind' as it develops from about two to four years of age that illustrate each of these properties. They conclude that therefore what the child has acquired over this time really *is* a theory because these properties of a scientist's knowledge are also properties of a child's 'theory of mind' knowledge.

The case of language

Unfortunately, the properties that Gopnik and Wellman (also Gopnik and Meltzoff, 1997) consider crucial to establishing their claim fail to distinguish knowledge entities that are indisputably real theories from knowledge entities that are merely 'theory-like.' Consider the case of language. The left-hand panel of Figure 2 mentions an indisputably *real* theory of language, namely, the *Principles and Parameters* theory of generative linguistics (e.g., Chomsky and Lasnik, 1995). This theory is widely regarded as being a piece of *echt* science even by those who do not regard it as being true. Furthermore, it is undoubtedly the case that some people (certainly not me) possess real knowledge of this theory. So here is a clear sense in which someone (e.g., Noam Chomsky) knows something and the something that he knows really *is* a theory.

The right-hand panel of Figure 2, by contrast, shows the psychological entities and mechanisms that (are postulated by the theory on the left to) embody the knowledge of language that people routinely possess,

[3] For critical discussion of this idea see Carey and Spelke (1996), Leslie and German (1995), and Stich and Nichols (1998).

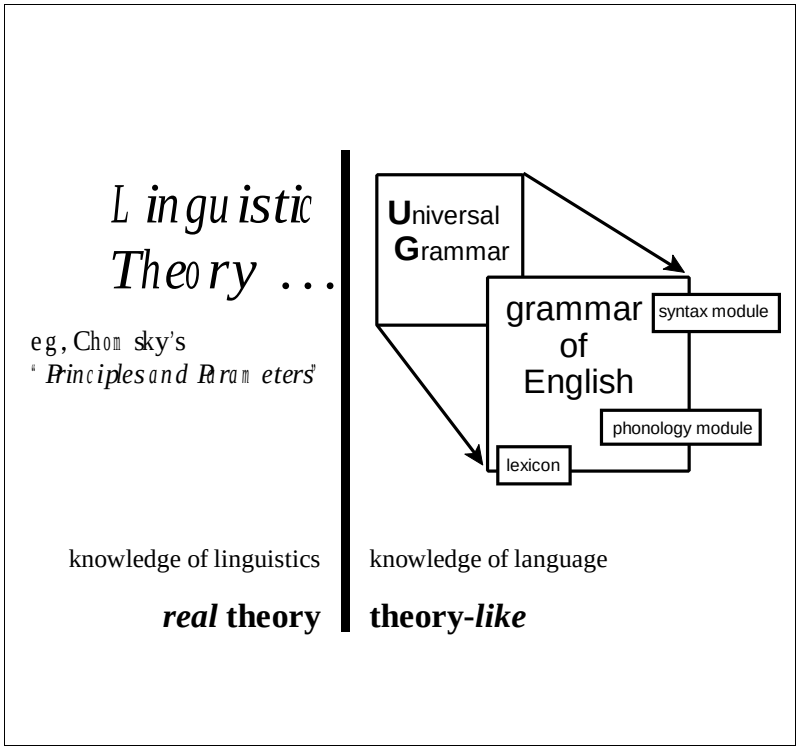

Figure 2 The case of language illustrates the distinction between a scientific theory ('real theory') and psychological entities that are theory-like. Both can be represented in the brains of people who possess the relevant knowledge: knowledge of linguistics and knowledge of language, respectively. However, most people only have knowledge of language.

including regular people like me and my neighbor's child, and not just special people like Chomsky. One of these entities is the "grammar of English," in some way represented in my brain and in the brain of my neighbor's child. Another entity is "Universal Grammar" which, according to the theory in the left-hand panel, is the entity, again in some way represented in the brain, that enabled me and my neighbor's child to acquire our knowledge of the "grammar of English." Chomsky's brain in some way represents all the entities depicted in Figure 2.

Now, a mental grammar has often been described as an internalization of a theory of a language, and the child's acquisition of a language has often been described as being like a process of theory formation, e.g., "[in acquiring knowledge of a language] the young child has succeeded in carrying out what from the formal point of view . . . seems to be a remarkable type of theory construction" (Chomsky, 1957:56). The entities or processes on the right of Figure 2 can reasonably be described as 'theory-*like*.' However, one would have to be *completely* blind to questions of mechanism to say that an internalized grammar, along with Chomsky's *Principles and Parameters*, really *is* a theory. Although almost nothing is known about the psychological basis of scientific knowledge, the best guess is that the child's knowledge of language is distinct from Chomsky's knowledge of linguistic theory in just about every respect that a psychologist might be interested in, including the mental representations involved, accessibility, penetrability, the timing, time course, and process of acquisition, and the underlying brain systems. Such distinctions are missed if we say that both knowledge of linguistics and knowledge of language really *are* theories.

As noted earlier, Gopnik and Wellman (1994, 1995 and also Gopnik & Meltzoff, 1997) argue that the child's 'theory of mind' really *is* a theory because it meets a set of criteria derived from a characterization of real theories. Unfortunately, these criteria also characterize the theory-like entities in the right panel of Figure 2 every bit as well as they characterize the *real* theory in the left panel. Theories postulate abstract entities that explain phenomena (Gopnik & Wellman, 1995:260): the child's internalized grammar is thought to 'postulate' abstract entities, e.g., categories like S and NP, properties of parse-tree geometry, and so forth, that explain sentence structure. Theories exhibit coherence in a system of laws or structures (Gopnik & Wellman, 1995:260): the child's internalized grammar is thought to be a system of interacting rules and representations that generate the structures of his or her language ('systematicity'). Theories make predictions "about a wide variety of evidence, including evidence that played no role in the theory's initial construction" Gopnik & Wellman, 1995:261): an internalized grammar allows the child to produce and comprehend novel sentences that "played no role in the [grammar's] initial construction" ('productivity'). Theories can be falsified by their predictions, yet may be resistant to counter-evidence, may spawn auxiliary hypotheses, etc. (Gopnik & Wellman, 1995:262-3): such phenomena in relation to the construction of an internalized grammar are much discussed in the language acquisition literature. Theories "produce interpretations of evidence, not simply descriptions . . . of evidence" (Gopnik & Wellman, 1995:262): internalized grammars produce interpretations of sound patterns in terms of

meaning via intermediate levels of structure including phonology, morphology and syntax, and not simply descriptions of the sounds themselves. Finally, although a "distinctive pattern of explanation, prediction and interpretation" such as we have noted above for grammar "is among the best indicators of a theoretical structure" (Gopnik & Wellman, 1995:262), it cannot distinguish a child's knowledge of language from Chomsky's knowledge of linguistic theory.

Modules and theory-theory

Gopnik and Wellman are not unaware that their criteria of 'theory-hood' are too weak to do much work. In contrasting their theory-theory view with the "innate module view" of the child's 'theory of mind', they note,

> ". . . many kinds of evidence that are commonly adduced to support [theory-theory] or [modularity], in fact, cannot discriminate between the two. . . the fact that the representations in question are abstract, and removed from the evidence of actual experience is compatible with either view." Gopnik & Wellman, 1994:282.

The failure to identify a *formal* basis for distinguishing between 'theory-like' knowledge structures (such as might be found in modular systems) and knowledge of 'real theories' should not be surprising. The philosophical project to develop a formal theory of what makes a set of beliefs into a scientific theory has long been abandoned as hopeless, as Gopnik and Wellman are aware. Many sets of 'beliefs,' even the 'beliefs' of perceptual systems, are abstract, coherent, predictive, explanatory, and offer interpretations that go beyond the evidence. There is no great harm in calling these systems 'theories' or 'theory-like.' But it is hard to see what the point might be in arguing that these systems 'really *are* theories' unless there's some definite way to distinguish them from systems which 'really *aren't* theories' but which are merely theory-like.

Gopnik and Wellman (1994, see also Gopnik and Meltzoff, 1997) advance one property of theories that they say discriminate theories from modules, namely, 'defeasibility.' The notion of defeasibility in the philosophy of science refers to the willingness of a theorist to regard a proposition or theory as 'negotiable' or revisable, for example, in the light of evidence. According to Gopnik and Wellman, this property of real theories is also a property of the commonsense theories that they attribute to children. Presumably, what they mean is simply that children's 'real theories' are revisable rather than that children always *believe* that their theories are revisable. In any case, according to these authors, modules are not similarly 'defeasible.' In fact, Gopnik and Wellman go so far as to label modules 'anti-developmental' (1994:283), apparently because they believe that knowledge in modules cannot be revised. They are careful to point out that it is not the issue of innateness that divides theory-theory from modularity theory. Indeed, they hold that theory-theory needs to postulate innate theories, including in particular, an innate 'theory of mind.' But these innate theories are not fixed for all time; they are 'defeasible' and are often quickly revised by the child.

However, even the property of 'defeasibility' does not discriminate between 'real theories' and 'theory-like' entities such as modules (see Stich & Nichols, 1998). It is hard to know why Gopnik and colleagues have come to believe that modules are fixed at birth, unrevisable, and 'anti-developmental.' None of the major modularity theorists posit such properties. Take the Chomskean modules of Figure 2 (right panel) as an example. The Universal Grammar module has the job of 'revising' itself in the light of the properties of the language(s) to which it is exposed. It does this by setting the values of a number of parameters. This in turn affects the nature of the grammar module that is constructed for a particular language. These modules learn and in the process 'revise' themselves and no doubt will have mechanisms to recover from error. My point is not that Chomsky's proposal is correct, just that in proposing modular processes Chomsky did not somehow overlook the fact that his modules were learning mechanisms. On the contrary, for Chomsky, that was the whole point. To take a rather different example of a module, consider Marr's (1982) 'Object Catalogue' whose job is to recognize 3-D objects from arbitrary viewing points. A module that performs this job has to learn the 3-D shapes of literally tens of thousands of everyday objects and no doubt makes the occasional error-plus-revision along the way. Again, my point is not that Marr's theory is right, just that in making his proposal, Marr, as an important modularity theorist, was quite happy that his module could perform a prodigious feat of learning. Or once again, consider the lexicon which modularity theorists, like Fodor (1983), often assume is a module. Given that the adult lexicon contains many tens of thousands of items (Levelt, 1999) and that infant lexicons contain none, the lexicon must learn on a grand scale, with the occasional recovery from error (Carey, 1978).

Innate theories and general learning

Gopnik and colleagues claim that modules are 'anti-developmental.' Perhaps they mean that the degree of defeasibility is too low, that 'theories' can be *radically* revised while modules can't. Wellman (1990) argues that the child's initial theory of belief is that "beliefs are copies of reality" but that this theory is soon revised to become the theory that "beliefs are representations of reality." Perhaps this is an example of radical revision of which modules are supposed incapable. The issues here are far from clear. However, it does seem odd that children should have an innate theory that almost immediately requires 'radical' revision and indeed that receives such revising within a year or two. If the necessary revisions to the innate theory become obvious to the average child between two and four years of age after applying his limited reasoning abilities to the morsel of idiosyncratic experience available in that time, why, with its vast experiential resources of biological time and whole populations, were these revisions not glaringly obvious to the processes of evolution or whatever Gopnik and colleagues assume bestowed the innate theory? Why doesn't Nature just bestow the revised 'theory' and be done with it? These are interesting questions, but, as Scholl and Leslie (1999b) point out, there is no reason to suppose that early 'theory of mind' involves 'radical revision' rather than plain learning. It is obvious why Nature should bestow a module that will contain more information at the end of its life than it does at the start. However, it is far from clear how the 'representational theory of belief' contains *more* information than the 'copy theory of belief,' rather than simply being a 'better' theory. And it is quite puzzling why Nature should bestow a *false* theory when she could have bestowed a true theory.

Perhaps what Gopnik and colleagues really want to say about theories versus modules is that theories are acquired by mechanisms of general learning whereas modules are mechanisms of specialized learning. Thus, someone acquiring knowledge of Chomsky's linguistic theories would have to employ mechanisms of general learning. Meanwhile, (according to Chomsky's theory) a child acquiring 'knowledge of language' employs specialized modular learning mechanisms. There are many interesting issues here that would take us too far afield to pursue. However, the evidence with regard to purely general mechanisms in 'theory of mind' development does not look good. Which general learning mechanisms might be involved? Presumably, exactly those that are used in scientific theory building. If that claim seems too strong, we can weaken it: if not those responsible for scientific creativity, then the mechanisms involved are those mechanisms involved at least in learning about scientific theories, or, at the very least, those involved in learning about 'science' at elementary levels of education. These mechanisms for 'real' science learning are highly sensitive to IQ, meaning that we find large differences between individuals in their ability to benefit from science education. Indeed, IQ tests were specifically designed to measure such differences in general or 'academic' intellectual ability (Anderson,1992). Mildly retarded individuals— for example, those with IQ's around 64 — have an extremely limited ability to acquire even elementary scientific ideas. Yet, mildly retarded non-autistic individuals can pass standard false belief tasks (e.g., Baron-Cohen et al., 1985; Happé, 1995). It has been clear for some time, then, that 'theory of mind' development is substantially independent of intellectual level and therefore cannot depend solely upon general purpose learning mechanisms. More recent evidence, some of it from unexpected sources, has also supported the modular nature of 'theory of mind' (Langdon & Coltheart, 1999; Leslie, in press; Varley & Siegal, in press).

Before I leave the question, I want to remark upon one property that real theories always have. It is impossible to imagine a scientific theory that is not explicitly articulated in a natural or a formal language. For example, Chomsky's knowledge of *Principles and Parameters* theory is explicitly articulated in a number of books and articles. Anyone who claims knowledge of Chomsky's theory must also be able to explicitly formulate its propositions, and to the extent he or she cannot do this, we deny them that knowledge. Translating this property into the 'real theory-theory' framework, we should say that knowledge cannot really *be* a theory unless it is explicitly articulated in a declarative representation. This places a strong requirement upon knowledge that is to count as a 'real theory:' it demands that the child be able to articulate, for example, his theory of belief. Is this too strong a requirement to place upon knowledge of a theory? It is if we want to allow 'implicit' knowledge of theories. Now, I am all in favor of implicit knowledge in theory-*like* entities and of leaving open to empirical investigation the question of which properties of a psychological entity are theory-like and which are not. That's the point of using *metaphors*. But can Gopnik and colleagues claim that a psychological entity really, non-metaphorically, *is* a theory and then get to pick and choose the properties in respect of which this is alleged to be true? Although I don't think they can, I shall put aside my misgivings. I shall not insist that the child be able to state (even) his 'real' theories.

However, I *will* insist that the theory-theorist be able

to articulate the child's theory — as it were, on the child's behalf. The articulable content of the child's theory forms the central substance of the claim made by the theory-theorist. In the case of Gopnik and colleagues, it is hard to discern exactly what the child's theory of belief is: What is it the child 'thinks' when the child entertains his 'representational theory of belief?' Surely, the child's theory can't simply be, "beliefs are representations." Why would *that* really *be* a theory? Both Gopnik and Wellman focus on what the younger child does *not* understand, but say little to specify what the older child's view actually is. Among the theory-theorists, only Perner has addressed this important point. I discuss Perner's specific proposals in the next section, after I have outlined the third and most interesting strand of current theory-theory. This version uses a theory analogy to provide an account of the semantics of abstract concepts.

Concept as theory

From this point on in the discussion, we will no longer worry about whether a 'theory' the child might have really *is* a theory. We will be content merely if a piece of knowledge is theory-like. In this section, we will be concerned principally with Perner's proposal, and Perner is not, as far as I know, committed to the 'theory of mind' really *being* a theory, in the sense of Gopnik and her colleagues. Perner (1991, 1995) is, however, committed to the child acquiring an explicit understanding of belief-as-representation, to the notion of conceptual change, and to the idea that "each particular mental concept gets its meaning not in isolation but only as an element within *an explanatory network of concepts*, that is, a theory" (Perner, 1991:109), and, therefore, to the idea of concept-as-theory.

The basic idea behind concept-as-theory is as follows. With something as abstract as *belief*, the only way that you could think thoughts about *belief*s is if you have a theory of what beliefs really are. Beliefs don't look like anything, they don't sound like anything, and they are not found in some specifiable location, and so forth, so how are you (your cognitive system/brain) going to describe (to yourself/itself) what a belief is? An attractive answer is that you will need something theory-like to specify what a belief is. The theory has to be accurate enough in its description of what a belief is to ensure that the concept, BELIEF, which is embedded in the theory, does in fact refer to beliefs and not to something else. The description is what will determine what is picked out by the concept. So, if the description does a very bad job (of describing what a belief is), and instead describes, say, a desire or a toothache, then the associated concept will not in fact be a concept of *belief* but a concept of *desire* or *toothache*, as the case may be. So the exact nature of the associated theory is vitally important because this is what determines both the *sense* of the concept and what its *referent* will be.

Moreover, on the concept-as-theory account, acquiring the concept, BELIEF, is acquiring the theory that says what kind of thing *belief* is. If the child has not acquired the theory, then he will not be in possession of the concept; if he acquires a theory that so badly describes *belief* that it instead describes *desire*, then the child will have acquired the concept DESIRE instead. It makes sense, then, on this version of theory-theory to pay a lot of attention to exactly what the child knows about *belief*. Because what he knows or doesn't know about *belief*, will determine what concept he has. To put it round the other way, you can discover what concept the child has by discovering what he knows or doesn't know about *belief*. But before you can decide whether the state of the child's knowledge means that he possesses the concept BELIEF, you must first decide what the critical knowledge is. This means you must decide what are *the* critical features of the adult concept BELIEF — what it is we big guys know about *belief* that makes our concept pick out just the things that are *belief*s. If you are a theory-theorist, this critical adult knowledge must be our commonsense theory of what *belief*s are. From the adult theory of *belief*, the developmental researcher derives a set of criteria that will be applied to the child's knowledge. If the child meets these criteria, he must possess the concept; if he does not, he must lack the concept. Hence the theory-theorist's interest in setting knowledge criteria for concept possession (Perner, 1991: Chapter 5).

The concept dictionary model

As I noted earlier, abstract concepts are widely supposed to be abbreviations for packets of knowledge. The concept-as-theory is one variant on this view. Imagine our repertoire of concepts as a dictionary — a long list of items, each made up of two parts: a concept on the left and an associated theory/definition on the right. Almost all the variance in theories of concepts has to do with the nature of the entries postulated for the right-hand side of the list: necessary and sufficient conditions (definitions), a stochastic function over features (prototypes), rules of inference, or theories. In every case, however, the entry on the right functions as some kind of a *description* of

whatever the concept on the left denotes. Hence the term Descriptivism for this general view of concepts. A dictionary model might be held explicitly, in the sense that its entries are assumed to be mental symbols or implicitly, in the sense that the entries are assumed to be merely emergent properties. Either way, *possessing* a given concept means having the correct entry for that concept in one's mental dictionary; *using* that concept (as the meaning of a word or as an element of a thought) is gaining access to the associated entry; and *acquiring* that concept means acquiring that entry.

Just as a real dictionary provides characterizations of words in terms of other words, so in the dictionary model of concepts it is assumed that the items on *both* the left and right sides of an entry are concepts. A concept is given a definition (or a prototype, theory, . . .) that itself is composed of concepts. For example, the entry for the concept DOG might give the definition, DOG = CANINE ANIMAL. In a prototype theory, DOG will be characterized as a stochastic function over properties such as HAIRY, FOUR LEGS, SLAVERS, BARKS, etc. A theory-theory might show an entry that makes critical reference to a dog being a LIVING THING. In all these cases, the descriptive entries are assumed to be made up of other concepts, such as CANINE, ANIMAL, HAIRY, LIVING THING, and so on, each of which will have its own entry with an associated description in the dictionary. That the descriptive entry is formed by other concepts is an especially natural assumption for the theory-theory, because it is hard to imagine how a theory could ever be stated without using concepts. In all dictionary model accounts, but in 'theory-theory' accounts in particular, possessing, using, and acquiring one concept depends upon possessing, using, and acquiring other concepts.

The dictionary model has a number of attractive features but it has one major drawback. The everyday word dictionary depends upon the fact that its user already knows the meanings of most of the words in the dictionary. If this wasn't true, the practice of defining one word in terms of a lot of other words would get nowhere. A dictionary in an utterly foreign tongue offers no point of entry or exit. If we know none of them, we can never escape from the maze of words and the dictionary is useless. The same point applies to the dictionary model of concepts. If we come to know what a given concept is by learning its (theoretical . . .) definition, which is given in terms of a lot of other concepts, then we will need already to possess those other concepts and already be able to pick out the things in the world to which they refer. But those other concepts are known by way of *their* entries in the concept dictionary which are comprised of a lot of still other concepts, and . . . Because this cannot literally go on forever, there must be some concepts which are known, not by a defining entry in the dictionary, but by some other route. These are usually called the *primitive* concepts. Primitive concepts provide the floor or ground upon which all other concepts are ultimately defined. A primitive concept is not acquired by learning a description; otherwise we are back in the maze. But, if there is a way to acquire a concept *without* learning a description, then the whole dictionary model is called into question. For this reason, dictionary models assume that primitive concepts are unlearned, i.e., innate.

With a highly abstract concept like BELIEF, the dictionary model creates a dilemma for theory-theory. Either BELIEF is primitive and innate, or it is acquired. If it is innate, then either the concept is constituted by an associated theory or it is not. If BELIEF *is* established by an associated theory (and *is* innate), then knowledge of that theory too *must* be innate. If it is *not* so constituted, then BELIEF is an abstract concept that falls outside the scope of theory-theory. And now we should ask for which other 'theory of mind' concepts theory-theory is irrelevant.

Alternatively, if BELIEF is acquired, then we have to ask: What are the *other* concepts, the ones in the associated description/theory/dictionary entry that the child has to acquire in order to possess BELIEF? Once we have an answer to that, we will be obliged to ask the same question about each of *those* concepts: What are their associated theories? What are the concepts in *those* theories? We must press our inquiries until, finally, we get answers that contain only primitive concepts. When we reach the innate primitive concepts, each of those concepts will either fall outside the scope of theory-theory or be constituted by an associated innate theory.

We can now understand the dilemma that BELIEF creates for theory-theory. When we pursue our repeated rounds of asking which concepts make up the associated theory that establishes BELIEF, the answers can go in one of two directions. Either the concepts in the associated entries become *less* abstract than BELIEF, or they become *more* abstract. If we assume they should be less abstract, we will end up characterizing BELIEF in behavioral terms. Theory-theorists correctly want to account for the *mentalistic* character of 'theory of mind' concepts but cannot do this by claiming that children are behaviorists. Alternatively, if we assume that the concepts in the associated entry for BELIEF are more abstract than BELIEF, we will find that our account ends up chasing larger and larger numbers of more and more abstract concepts, most of them quite obscure, while the possibility of accounting for their acquisition slips further and further from our grasp.

A close up of the representational theory-theory

Perner (1988, 1991) proposed that the four-year-old child comes to pass false belief tasks by discovering the representational theory of mind and in particular the representational theory of belief. Younger children adhere to a different theory, namely, that people are 'mentally connected' to *situations*, a theory which is meant to preclude conceptualizing belief such that the content of a belief can be false. Older children then make a theoretical advance, discovering that beliefs are really representations; this advance creates a new concept, namely, BELIEF, and ushers in success on false belief tasks.

When Perner originally proposed the representation theory-theory, the idea was that the child discovered that mental states were like *other* representations — like pictures or models, for example. Perner wrote,

> "If we define representation . . . as I have done, then we use the word "representation" to refer to the representational medium (more precisely the state of the medium). For instance, in the case of a picture it is the picture (medium) that is the representation and not the scene depicted on it (content)" (1991:280).

The key development in the child's 'theory of mind' was then said to occur around four years when the child acquired the (supposedly adult-like and commonsense) theory that mental states are internal representations. This, in turn, was said to be achieved by the child coming to "model models" by "work[ing] out the notion that something (referent) is apprehended (represented) as something (sense)." (Perner, 1991:284).

As Leslie and Thaiss (1992) point out, the most natural supposition for a representational theory of mind is that children acquire a representational theory of belief by hypothesizing that beliefs are internal mental pictures. Sally puts the marble in her basket and makes a mental picture or takes a mental photograph of the marble in the basket. Then she goes away with her mental picture. While she is away, naughty Ann discovers the marble and moves it from the basket to the box. Now, Sally is coming back! Where will she look for her marble? Answer: Sally will consult her mental picture which will show her that the marble is in the basket. This idea is highly attractive for a number of reasons. First, it provides a series of thoughts that preschool children might actually have, avoiding obscure and ultra-abstract concepts. Secondly, it would explain how preschoolers come to have the concept BELIEF by learning about things, like pictures, that are visible, concrete objects rather than invisible 'theoretical' constructs. Thirdly, mother can show you pictures, she can point to them, count them, discuss and compare them with you; in short, she can tutor you about pictures in ways she cannot tutor you about beliefs. Finally, almost every picture or photograph you have ever seen is 'false' or out-of-date, making them ideal for learning about their representational proprieties — about how something (you, a *big* boy or girl) is represented as something else (a baby).

Coming to solve the false belief task by way of a picture theory implies that understanding an out-of-date picture is a sub-component of understanding an out-of-date belief. If a picture task is a component task, then it cannot possibly be harder than a false belief task and, if anything, ought to be easier. Using tasks adapted from Zaitchik (1990), Leslie and Thaiss (1992) showed that out-of-date pictures are not easier and, in fact, are slightly harder, at least for normally developing children. For children with autism, Leslie and Thaiss showed exactly the opposite is true (see also Charman & Baron-Cohen, 1992, 1995). Understanding out-of-date pictures is therefore neither a necessary nor a sufficient condition for passing a false belief task. These findings are a blow to the idea that the child "works out" that beliefs have a "representational medium" (for further discussion, see Leslie and Thaiss, 1992, Leslie and Roth, 1993, and Leslie, 1994).

In light of these sorts of findings, Perner (1995) abandoned his original version of representational theory-theory. Rather than having to master a *general* theory of representation, the child is now said to employ a theory of representation *specific* to understanding beliefs.[4]

---

[4] Slaughter (1998) claims that the dissociation between children's performance on false belief tasks and photographs tasks is predicted by Gopnik and Wellman's theory-theory on the grounds that "[a]lthough theory-building processes require general cognitive skills and resources, the resultant concepts, including mental representation, are held to be specific to the domain of folk psychology" (p330). It is hard to see what property of Gopnik and Wellman's views predicts that concepts/theories should be specific in this way. Certainly, the opposite is true of real theories which strive for as much generality as possible. Indeed, the representational theory of mind is exactly the attempt to treat mental states as instances of something more general, viz., as representations. Without this generality, it is not obvious even what is meant by 'representational' in the phrase 'representational theory of mind.'

(continued...)

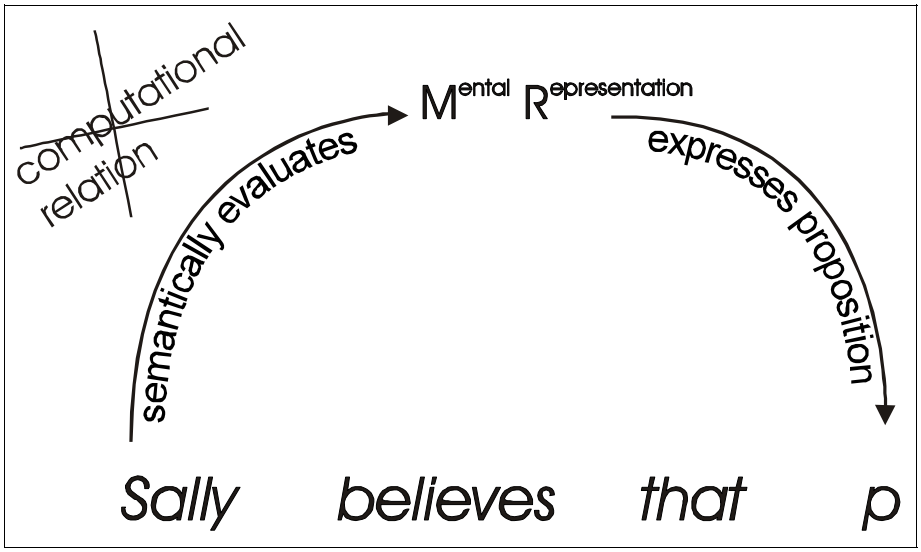

Figure 3: Perner's latest proposal borrows from cognitive science the idea that a belief is a relation to a mental representation. But instead of referring to a computational relation, the preschool child's BELIEF critical knowledge refers to a semantic evaluation relation to the mental representation. If Gopnik views the child as 'little scientist,' Perner views the child as 'little Fodor.' (After Perner, 1995.)

In characterizing the new theory-theory, Perner draws upon Fodor's explication of the theoretical foundations of cognitive science. Fodor (1976, 1981a) argues that a propositional attitude, such as *believing that p*, should be understood as a *computational relation* between an organism and a mental representation expressing the proposition *p*. Fodor's account is intended as a scientific account of what propositional attitudes really are. Perner attributes knowledge of this account to the child with one modification: instead of the child conceptualizing the notion COMPUTATIONAL RELATION, Perner says that the preschooler uses the concept SEMANTICALLY EVALUATES. According to Perner (1995), in order to understand that *Sally believes that p*, (in the case that *p* is false), the child must construct the 'metarepresentation,' *Sally semantically evaluates a mental representation expressing the proposition that p* (see Figure 3).

The first thing to notice is that moving from a general theory of representation to a specific theory of mental representation deprives the theory of any independent evidence. The 1991 version could hope to draw upon independent evidence that children first understand the idea of representing something as something else in regard to external, public representations like pictures, maps or models and then project these ideas to internal mental states. But, as we saw above, such independent evidence has evaporated. This has the disturbing consequence that the evidence supporting the idea that the child can understand that *Sally semantically evaluates a mental representation expressing the proposition that p* is just the evidence that supports the idea that the child can understand *Sally believes that p*, namely, passing false belief tasks. Therefore, there is, at present, no (independent) evidence to support the new theory-theory.

Let us remind ourselves of how Perner got to this position. He accepts the theory-theory account of concept possession: to possess the abstract concept BELIEF is to possess critical knowledge about *belief* and to acquire the concept is to acquire the critical knowledge. The critical knowledge in question is a theory of what *belief* is. In order to state the theory of *belief*, other concepts must be used. Therefore, the possessor of BELIEF must also possess these other concepts (the ones used to state the theory of *belief).* Rightly, Perner eschews the constraint that these other concepts must be *less* abstract than BELIEF. If, ultimately, BELIEF can be cashed out as or reduced to sensory concepts, then theory-theory is not really required. Moreover, reduction would entail that the child's (and our adult) 'theory of mind' concepts are fundamentally behavioristic and non-intentional. Rightly though, theory-theory is committed to mentalism. But rejecting this route, Perner is forced to allow the theory-

[4] (...continued)

Incidently, Slaughter (1998) overlooks the fact that in her study she compared children's performance on a modified photographs task with an unmodified false belief task. Just as it is possible to modify false belief tasks to make them easier for three-year-olds to pass, so it should be possible to modify photographs tasks too. Slaughter's results confirm this. According to the Leslie and Thaiss (1992) model, in making the comparison she did, Slaughter removed the only limiting factor that photograph tasks and false belief tasks have in common, namely, "selection processing." The resulting lack of correlation in children's performance does not "call into question the . . . model of development offered by Leslie" as Slaughter claims, but actually supports the model.

explicating concepts to be more abstract.

Perner is also forced to choose a theory of *belief* that might plausibly be true. The theory of *belief* that he requires has to explain how a thought containing the concept BELIEF actually picks out *belief* rather than something else, such as desire, serious facial expressions, an earnest gesture, or some other property of a situation containing a person with a belief. If the child (e.g., the three-year-old) has the wrong theory, then *his* concept BELIEF* will pick out something different from *our* concept BELIEF. And what theory can do the job of picking out *belief* other than *our* theory of what a *belief* really is?

However, there is a heavy price for taking this approach. In order to discover and apply the above representational theory of belief, the child must acquire the following concepts: SEMANTIC, EVALUATE, MENTAL, REPRESENTATION, EXPRESS, and PROPOSITION. The child *must* acquire these concepts because these are the concepts that state the critical theory of *belief*. Therefore, the child couldn't understand this theory unless he or she grasped these concepts. And if the child didn't understand this theory, then, according to Perner and theory-theory, the child wouldn't possess the concept BELIEVES.

We began by asking how one 'difficult' and obscure concept is acquired (BELIEVES), but now we have six more, each of which is just as 'difficult' and considerably more obscure. It is every bit as puzzling how the child might acquire any one of these six notions as it is puzzling how he acquires BELIEVES. One answer might be that these six concepts are innate. But if we are willing to accept *that*, why weren't we willing to accept that BELIEVES is innate? If we are not willing to accept these 'new' concepts as innate primitives, then each must, like BELIEVES, be acquired by acquiring and possessing critical knowledge — i.e., by acquiring a theory of *semantic evaluation*, a theory of *mental*, a theory of *representation*, and so on. Each of these theories will spin off further abstract concept-theory cycles, with no obvious end in sight. If we balk at *this* point in pursuing a theory-theory of concept possession and acquisition, the question inevitably arises why we didn't balk earlier at the first step: the decision to pursue a theory-theory of BELIEVES.

Unfortunately, the situation for the 'mental representation' theory-theory of *belief* is even worse than we have suggested so far. Fodor's formulation of propositional attitudes as computational relations to mental representations was designed to say what propositional attitudes *in general* are. It was not designed to characterize specifically *beliefs*. Fodor's formulation therefore does not distinguish *beliefs* from other mental states, such as *desires*, *hopes*, *pretends*, and so forth — they are *all* computational relations to mental representations. Each different attitude is assumed to involve a different kind of computational relation, putting it on the agenda of cognitive science to develop theories of each of the specific computational relations involved. This general characterization of propositional attitudes carries over into Perner's replacement of *computational relation* by a *semantic evaluation* relation (Figure 3). All propositional attitudes 'semantically evaluate' their 'mental representations'— by definition, propositional attitudes are attitudes to the truth of a proposition. So, even if the child *did* discover this obscure theory, it would still not provide him or her with the concept BELIEF, but only with an undifferentiated concept of *propositional attitude*. The theory in Figure 3 will only tell the child about propositional attitudes *in general*, applying to *desires* and *pretends* equally as it applies to *beliefs*. It will even apply just as well to *'prelief,'* the pretend-belief state that Perner, Baker and Hutton (1994) suggest three-year-olds attribute to other people. What it will *not* do is tell the child specifically what a *belief* is.[5]

Can the theory-theory in Figure 3 be patched up so that it provides to the child a theory of what *belief* is (as

[5] Fodor (pers. com.) points out that *believing that p* cannot be the same thing as *evaluating or holding-true a representation that means that p*. Consider: I have in my hands a copy of Einstein's paper on Special Relativity. I have never read this paper and, to be honest, I don't have a clue what it says. However, I know that the theory expressed in this paper is a cornerstone of modern physics, which, as far as I'm concerned, means that it's true. Secondly, this bunch of paper I have in my hands is only a representation of Einstein's theory. So I semantically evaluate (as true) this representation expressing the Special Relativity Theory. However, there is not a single proposition expressed in this paper that I have as a belief in the usual sense because I have no idea which propositions this paper expresses. But *whatever* they are, I hold them all to be true because I trust physicists to know what's what. But when I think that Sally believes that *p*, the marble is in the basket, I think that she actually grasps that very proposition. The idea behind treating belief as a particular kind of computational relation is that an organism standing in such a relation will *thereby* grasp and believe the proposition expressed. Without that assumption, a computational account of belief will not work. However, as the above example shows, exactly this assumption fails for the semantic evaluation relation. This is a further reason why substituting *semantic evaluation* for *computational relation* will not provide a theory of *belief*.

opposed to *desire*, *pretense*, etc.)? The main problem is with the relation 'semantically evaluates.' 'Mental representations that express propositions' will be a common feature in theories of *belief*, *desire*, *pretense*, *hopes*, etc. because Sally can desire that her marble be in the basket, can pretend that her marble is in the basket, or hope that her marble is in the basket, as well as believe that's where it is (and all the while the marble is in the box). What differs from case to case is the 'mode' of evaluation. The obvious temptation then is to add a simple qualification: Sally 'semantically evaluates *with respect to believing*' a mental representation . . . Certainly, this will do the job; but so will simply replacing 'semantically evaluates' with 'believes.' And both will work for exactly the same reason: namely, the concept BELIEF has been smuggled in. But this makes the theory-theory circular. *Belief* certainly is *belief* and, yes, the child will acquire the concept BELIEF by acquiring the concept BELIEF. But, theory-theories are simply not allowed to say that!

Agenda for a successful theory-theory

Here is a minimal agenda for a theory-theory of BELIEF. The first problem is to say, *without circularity*, what belief really is. Having made explicit the theory that constitutes the critical knowledge for concept possession, the next step is to provide *independent* evidence that the child does in fact acquire this critical knowledge. Finally, it must be shown that it is *by* acquiring this critical knowledge that the child acquires the target concept. Present accounts fall far short of achieving any of these goals.

Considerable obstacles lie in the way. I identified three sets of problems that face theory-theories of belief, and probably theory-theories more generally. The first is the *conceptual explosion* caused by concepts having 'dictionary entries' — that is, theories attached to each concept that have to say what sort of thing the referent of the concept really is, in order that the concept picks out that referent. Because theories are themselves composed of concepts, the number of concepts for which the theory-theorist must seek an acquisition account grows explosively. Proposing 'conceptual holism' sounds fine but I suspect that in practice it is quite frustrating. Secondly, because theory-theorists reject the idea that all abstract concepts reduce to statements formed solely of sensory concepts (you certainly can't be a theory-theorist if you accept that doctrine), the concept explosion will involve the *escalating obscurity* of the concepts that are spun off (cf. Fodor, 1981b). Perner's proposals illustrate this nicely: on the first iteration alone, we move from worrying about BELIEF to worrying about SEMANTIC. And *what* is the theory, grasped by the child, that constitutes the concept SEMANTIC? I suspect that escalating obscurity is a general feature of theory-theories — cf., DADDY = MALE REPRODUCER. Finally, the critical knowledge for BELIEF, conceptually rich and potent though it was in Perner's proposal, still fell short of the mark in specifying the exact meaning of BELIEF. This means that such a concept specification would not in fact do its critical job of specifically picking out *belief*s but instead would pick out any and all propositional attitudes. I suspect that the search for critical knowledge will only provide a *paraphrastic approximation* that forever falls short of its target — unless one introduces circularity.

Whether theory-theory can overcome these obstacles remains to be seen. In the meantime, alternative approaches should be vigorously explored.

Concept as soap molecule

One avenue to explore is dropping the notion that the *sense* of a concept — its associated critical knowledge — determines its reference and, therefore, which concept it is. The reference of a concept must be determined some how but stored knowledge is not the only conceivable way. In fact, it is far from clear how sense is supposed to determine reference. How does a concept fit to the world? Answer: A concept points to a stored description (of some sort); the description is laid against the world by the cognitive system and the things that fit the description are admitted to membership of the set of things in the concept category. But if a description is itself composed of concepts, saying that a concept fits to the world via a description does not answer the question of how a concept fits to the world. It provides a postponement instead of an answer.

Historically, there used to be a non-question begging answer. Empiricist philosophers, like Hume (1740), argued that concepts fall into two major types: the sensory and the abstract (in his terminology, "impressions" and "ideas," respectively). Possession of the sensory concepts is provided directly by the structure of the sensory apparatus. That is, Hume assumed that the way a concept like RED locked to the world did not entail applying knowledge of a description of *redness* (whatever that would be). The locking was provided by a mechanism, namely, the mechanisms of color vision. Such mechanisms can be innate and provide an innate concept RED without us having to suppose that *therefore* some piece of knowledge or theory about *redness* is innate. An

infant doesn't need a theory of *redness* if she possesses something else, namely, the mechanisms of color vision. Possession of the sensory concepts does not require knowing a dictionary definition or theory because these concepts are locked to the appropriate property in the world by sensory mechanisms.

So long as all non-sensory (abstract) concepts reduce to descriptions composed entirely of sensory concepts, we have a general outline for abstract concepts of how their sense determines their reference. But without that reductionist assumption about abstract concepts, we lack a proposal for how sense could determine reference. Theory-theory rightly rejects the notion that all abstract concepts reduce to sensory descriptions. But as we saw, this raises a number of obstacles to understanding how certain abstract concepts can appear so early in life. These problems could be avoided if there was some way for an abstract concept to lock to the world, other than through applying critical knowledge. In the case of sensory concepts, such an alternative has seemed uncontroversial: a sensory concept is locked to target by a psychophysical mechanism. Can this idea be extended to abstract concepts too?

The idea that certain abstract concepts might be acquired by way of a mechanism that locks to a specific target property in the world is certainly a wild idea. But is it wild enough to be true? There is a philosophical tradition (that has received far less attention than the mainstream Descriptivist accounts to which theory-theory is heir) which has tried to develop causal theories of reference (e.g., Fodor, 1998; Kripke, 1972; Margolis, 1998; Putnam, 1975). The fundamental idea is that concepts bear information about a specific property, not because of subjective knowledge, but because of an entirely objective causal relation between the concept (as psychological entity) and a property ('in the world'). Such views stress the 'psycho-physical' duality of concepts. Like a soap molecule with one pole locked to oil and the other pole locked to water, a concept has one 'pole' locked to the causal processes of a cognitive system and the other 'pole' causally locked to the world. Instead of being lost in the endless maze of mutual inter-definition, the representational relation between concept and world is brought directly to the fore.

What is the role of knowledge in 'conceptual psychophysics?' Knowledge about the referent of a concept is acquired and associated with the concept, but this stored associated knowledge does not provide or constitute the sole locking mechanism for the concept. So the knowledge is free to change without affecting what the concept designates.

But isn't it true that we acquire new knowledge and that this new knowledge changes the way we think about something? Don't we learn about *beliefs* or *daddies* or *dogs* so that we come to see them in a 'new light?' Most certainly we do. What is at stake is not *whether* we learn, or whether that learning leads us to 'conceive' of things in a new way. What is at stake is whether, once our concept DADDY is locked to the world, its reference changes systematically in relation to our evolving ideas about what a *daddy* really is. According to the Descriptivist view (and theory-theory), it does. According to conceptual psychophysics, it does not. We can capture our strong intuition about changes in the way we 'conceive' of things by distinguishing between *concepts* and *conceptions*. 'Concept' will refer strictly to the symbol *cum* reference-relation-to-a-property, while 'conception' will refer to any knowledge associated with the symbol. 'Conception' will capture what we know or believe about whatever the concept refers to. Since what we believe about something determines how it appears to us, we can retain the intuition that new knowledge changes how we think about things. What new knowledge will *not* do is change the meaning of our concepts.

In theory-theory, or any Descriptivist approach, the claim that a given (abstract) concept is innate, entails that critical knowledge is innate. In a conceptual psychophysics framework, this entailment does not hold. A concept may be innate if at least one locking mechanism is innate (there does not have to be a unique or 'critical' mechanism). The existence of innate knowledge remains an empirical question, of course, and it is even possible that innate knowledge may play a role in a given locking mechanism. Likewise, in theory-theory, or any Descriptivist approach, the *acquisition* of a given concept entails the acquisition of critical knowledge. Again, this entailment does not hold within a conceptual psychophysics approach. Acquiring a new concept will mean acquiring a lock on a new property.

It seems to me that a good way to study these questions empirically is concept by abstract concept. Although there are a great many concepts, it would be a great advance to have an account for even a single abstract concept of how it is innate or how it is acquired. There have already been suggestive findings. For example, Leslie and Keeble (1987) showed that six-month-old infants recognized a specifically causal property of events in which one object launched another by colliding with it. They proposed that infant recognition was based upon a modular mechanism operating independently of general knowledge and reasoning to "provide information about the spatiotemporal and causal structure of appropriate events" and that "it could do this without having to know what a cause 'really' is" (p.286). Such a mechanism

would allow the infant to attend to physical causation, to lock in the concept CAUSE, and then begin to learn about causal mechanisms from instances. There are also promising ideas concerning locking mechanisms for number concepts (see, e.g., Gallistel & Gelman, 1992) and faces (Johnson & Morton, 1991). Recently, Leslie, Xu, Tremoulet & Scholl (1998; see also Scholl & Leslie, 1999a) have suggested an account of how the infant's concept of *object* gets locked without recourse to knowledge of a theory of objecthood. Finally, in the 'theory of mind' domain, Leslie (1987) proposed a model of how the concept PRETEND is locked without assuming that the infant has critical knowledge of what pretending really is (see also German & Leslie, unpublished). In a similar vein, Leslie (in press) discusses the development of the concept BELIEF as part of a mechanism of selective attention.

So, how do you acquire a representational theory of mind?

In the theory-theory account, the child discovers a theory of general (or alternatively, mental) representation that gives birth to the concept BELIEF and to success on false belief problems. In this chapter, I have laid out a number of reasons that make me skeptical of this claim. In fact, I think the relationship between concept and theory is exactly the reverse. It is the possession of the concept BELIEF (plus a gradual increase in skill at employing the concept) that eventually gives rise to a commonsense representational theory of mind. As the child begins to enjoy increasing success at solving false belief problems, he or she will increasingly *notice* false beliefs and the circumstances that give rise to them. In an everyday sense, the child will then develop commonsense 'theories' about how other people represent the world. For example, if Mary represents bananas as telephones, the child can model this fact as *Mary thinks bananas are telephones*. Or if the child sees a dog chase a squirrel which then runs up tree B, while the dog goes barking up tree A, the child can 'theorize' that *the dog is barking up the wrong tree because it thinks there is a squirrel up there*. In the limited manner of commonsense theory and opinion, this is a representational theory of the dog's mind. If it is disappointing to find that the child's 'representational theory of mind' is so mundane and epiphenomenal on the child's concept of *belief*, at least we know that children actually think thoughts like these. There is no evidence that children *ever* explicitly think thoughts of the sort in Figure 3.

References

Anderson, M. (1992). *Intelligence and development: A cognitive theory*. Oxford: Blackwell.

Baron-Cohen, S., Leslie, A.M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, **21**, 37–46.

Boyer, P. (1994). Cognitive constrains on cultural representations: Natural ontologies and religious ideas. In L.A. Hirschfeld and S.A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*. (pp. 391–411). Cambridge, UK: Cambridge University Press.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G.A. Miller (Eds.), *Linguistic theory and psychological reality*, Cambridge, Mass.: MIT Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (1988). Conceptual differences between children and adults. *Mind & Language*, **3**, 167–181.

Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, **63**, 515–533.

Charman, T., & Baron-Cohen, S. (1992). Understanding drawings and beliefs: A further test of the metarepresentation theory of autism (Research Note). *Journal of Child Psychology and Psychiatry*, **33**, 1105–1112.

Charman, T., & Baron-Cohen, S. (1995). Understanding photos, models, and beliefs: A test of the modularity thesis of theory of mind. *Cognitive Development*, **10**, 287-298.

Chomsky, N.A. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N.A., & Lasnik, H. (1995). The theory of principles and parameters. In N.A. Chomsky, *The Minimalist Program*. (pp. 13–127). Cambridge, MA: MIT Press.

Flavell, J.H., Green, F.L., & Flavell, E.R. (1993). Children's understanding of the stream of consciousness. *Child Development*, **64**, 387–398.

Fodor, J.A. (1976). *The language of thought*. Hassocks, Sussex: Harvester Press.

Fodor, J.A. (1981a). Propositional attitudes. In J.A. Fodor (Ed.), *Representations: Philosophical essays on the foundations of cognitive science*. (pp.177–203). Brighton: Harvester Press.

Fodor, J.A. (1981b). The present status of the innateness controversy. In J.A. Fodor (Ed.), *Representations: Philosophical essays on the foundations of cognitive science*. (pp. 257–316.) Cambridge, MA: MIT Press.

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

Fodor, J.A. (1998). *Concepts: Where cognitive science went wrong*. Oxford: Clarendon Press.

Gallistel, C.R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, **44**, 43–74.

German, T.P., & Leslie, A.M. (unpublished). Children's inferences from *knowing* to *pretending* and *thinking*. MS. Rutgers University Center for Cognitive Science.

Gopnik, A., & Meltzoff, A.N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H.M. (1994). The theory theory. In L. Hirschfeld and S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*, (pp. 257–293). New York: Cambridge University Press.

Gopnik, A., & Wellman, H.M. (1995). Why the child's theory of mind really *is* a theory. In M. Davies & T. Stone, (Eds.), *Folk psychology: The theory of mind debate*. (pp. 232–258). Oxford: Blackwell.

Happé, F.G. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development*, **66**, 843–855.

Hume, D. (1740). *A treatise of human nature*. London: Clarendon, 1978.

Johnson, M.H., & Morton, J. (1991). *Biology and cognitive development.* Oxford: Blackwell.

Keil, F.C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kripke, S.A. (1972). Naming and necessity. In D. Davidson and G. Harman (Eds.), *Semantics of natural language*, (pp. 253–355). Dordrecht: Reidel.

Langdon, R., & Coltheart, M. (1999). Mentalising, schizotypy, and schizophrenia. *Cognition*, **71**, 43–71.

Laurence, S., & Margolis, E. (in press). Concepts and cognitive science. In E. Margolis and S. Laurence (Eds.), *Concepts: Core Readings*. Cambridge, MA: MIT Press.

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

Leslie, A.M. (1994). *Pretending* and *believing*: Issues in the theory of **ToMM**. *Cognition*, **50**, 211–238. Reprinted in J. Mehler and S. Franck (Eds.), *COGNITION on cognition*, pp. 193–220. (1995). Cambridge, MA.: MIT Press.

Leslie, A.M. (in press). 'Theory of mind' as a mechanism of selective attention. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences*, 2nd Edition, Cambridge, MA: MIT Press.

Leslie, A.M., & German, T.P. (1995). Knowledge and ability in “theory of mind”: One-eyed overview of a debate. In M. Davies and T. Stone (Eds.), *Mental simulation: Philosophical and psychological essays*. pp. 123–150. Oxford: Blackwell.

Leslie, A.M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, **25**, 265–288.

Leslie, A.M., & Roth, D. (1993). What autism teaches us about metarepresentation. In S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen (Eds.), *Understanding other minds: Perspectives from autism*, (pp. 83–111). Oxford: Oxford University Press.

Leslie, A.M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, **43**, 225–251.

Leslie, A.M., Xu, F., Tremoulet, P., & Scholl, B. (1998). Indexing and the object concept: Developing ‘what’ and ‘where’ systems. *Trends in Cognitive Sciences*, **2**, 10–18.

Levelt, W.J.M. (1999). Models of word production. *Trends in Cognitive Sciences*, **3**, 223–232.

Margolis, E. (1998). How to acquire a concept. *Mind & Language*, **13**, 347–369.

Marr, D. (1982). *Vision*. San Francisco: W.H.Freeman & Co.

Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review*, **92**, 289–316.

Perner, J. (1988). Developing semantics for theories of mind: From propositional attitudes to mental representation. In J. Astington, P.L. Harris and D. Olson (Eds.), *Developing theories of mind*, (pp. 141–172). Cambridge: Cambridge University Press.

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

Perner, J. (1995). The many faces of belief: Reflections on Fodor’s and the child’s theory of mind. *Cognition*, **57**, 241–269.

Perner, J., & Wimmer, H. (1988). Misinformation and unexpected change: Testing the development of epistemic state attribution. *Psychological Research*, **50**, 191–197.

Putnam, H. (1975). The meaning of 'meaning'. In K. Gunderson (Ed.), *Language, mind, and knowledge*, (pp. 131–193). Minneapolis: University of Minnesota Press .

Scholl, B.J., & Leslie, A.M. (1999a). Explaining the infant’s object concept: Beyond the perception/cognition dichotomy. In (Eds.), E. Lepore & Z. Pylyshyn, *What is Cognitive Science?* (pp. 26–73). Oxford: Blackwell.

Scholl, B.J., & Leslie, A.M. (1999b). Modularity, development and ‘theory of mind’. *Mind & Language*, **14**, 131–153.

Slaughter, V. (1998). Children’s understanding of pictorial and mental representations. *Child Development*, **69**, 321–332.

Stich, S., & Nichols, S. (1998). Theory-theory to the max: A critical notice of Gopnik & Meltzoff's *Words, Thoughts, and Theories*. *Mind & Language*, **13**, 421–449.

Varley, R., & Siegal, M. (in press). A dissociation between grammar and cognition. *Nature*.

Vosniadou, S. (1994). Universal and culture-specific properties of children’s mental models of the earth. In L.A. Hirschfeld and S.A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*. (pp. 412–430). Cambridge, UK: Cambridge University Press.

Wellman, H.M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, **13**, 103–128.

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and 'false' photographs. *Cognition*, **35**, 41–68.

To appear in D. Sperber and S. Davis (Eds.), *Metarepresentation. Vancouver Studies in Cognitive Science, Vol. 10*. Oxford: Oxford University Press

How to acquire a 'representational theory of mind'

Alan M. Leslie[1]

Department of Psychology and
Center for Cognitive Science
Rutgers University

The study of cognitive development is dominated by the view that concepts are essentially *packets of theory-like knowledge* (Carey, 1985, 1988; Keil, 1989). This view has emerged from a long tradition of viewing concepts as descriptions of one kind or another, though there have been and continue to be many variations and disagreements concerning the character of the associated knowledge (e.g., Murphy & Medin, 1985; for critical reviews of this tradition, see Fodor, 1998; Kripke, 1972; Laurence & Margolis, in press). The essence of this family of views is that the knowledge packet associated with the concept determines what in the world a given concept refers to or designates — it fixes what the concept is a concept of. For example, the concept DOG[2] might be associated with a knowledge structure that specifies "HAIRY, FOUR-LEGGED, ANIMAL, BARKS, WAGS TAIL, . . ." If this specification captures the structure of the concept, DOG, then this is the specification you will need to know in order to possess the concept, for the following reason: this specification, when applied to the world, is what will selectively pick out the things that are *dogs* and thus link DOG with *dogs*.

The 'concept as knowledge' view is so deeply entrenched that it is hard to see how there could be an alternative. The view has two powerful implications for conceptual development. First, the acquisition of a concept must be the acquisition of the critical knowledge that defines the concept. Second, the innateness of a concept must be the innateness of the critical knowledge that defines the concept.

Perhaps the knowledge view of concepts will prove to be correct. However, to date, there is not a single concept for which a detailed model of the critical knowledge has been worked out and empirically substantiated; there is not a single concept whose acquisition or innateness has been understood. All conclusions therefore remain highly tentative.

Much of the most interesting work in cognitive development over the last twenty years has been concerned with abstract concepts, that is, with concepts that are not reducible to sensory transduction. Many abstract concepts are now thought to emerge early in development. Mental state concepts, such as BELIEVE, DESIRE, and PRETEND, are among the most abstract we possess. It is striking that these concepts are routinely acquired by all normally developing children before they attend school and are even acquired by children who are mentally retarded. The verbal labels associated with these concepts are never explicitly taught, yet are typically in use around the third birthday; by contrast, words for colors, a salient sensory property, very often are explicitly taught by parents, but are typically not learned any earlier and are often learned later. Mental state concepts provide a crucial challenge to our attempts to understand

[1] *Acknowledgments*: I am grateful to the following friends and colleagues: Eric Margolis and Susan Carey for helpful discussions, and to Jerry Fodor, Shaun Nichols, and Brian Scholl for helpful discussions and detailed comments on an earlier draft.

[2] I use small caps when referring to a concept as opposed to what the concept denotes (italicized). Normally, one could simply say that the concept is a psychological entity, while what it denotes is not, e.g., DOG refers to *dogs*. But in the case of mental state concepts what they denote are also psychological entities.

what is required for the acquisition and possession of abstract concepts. In our attempts to understand early emergence, one variant of the knowledge view of concepts has become popular; in this variant, critical knowledge is said to take the form of a *theory*. The concept BELIEF has been a central focus of these attempts.

At first sight, it is plausible that the acquisition of the concept BELIEF must be theory formation because how else can we come to know abstract things, if not by employing theories. The so-called 'theory-theory' of BELIEF has gained a widespread credence (Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994, 1995; Perner, 1991; Wellman, 1990). However, I believe that current attempts to develop a theory-theory of BELIEF have foundered. In this chapter, I will explore the reasons for these present difficulties. Because I have elsewhere written extensively on the relevant experimental evidence and developed an alternative framework to theory-theory (see e.g., Leslie, in press for a review), here I shall confine myself to examining the deeper motivations for theory-theory in order to say why I believe the entire enterprise is mistaken.

Three versions of theory-theory

There seems to be about three versions of 'theory-theory' currently active; they are not always clearly distinguished, though they need to be. The first is simply the idea that not all knowledge is sensory in character and that some knowledge is concerned with 'understanding' the world. This seems sensible and true. To say that people acquire commonsense 'theories' in this sense is just to say that they acquire abstract knowledge and opinion. For example, people develop opinions about the existence of ghosts (Boyer, 1994), the nature of consciousness (Flavell, Green & Flavell, 1993), and the disposition of heavenly bodies (Vosniadou, 1994). People also develop opinions about circumstances which will cause beliefs to be false. This might be called a 'representational theory of mind' and, if so, I shall argue that the concept BELIEF is prior to the theory.

A second current version of theory-theory is more controversial. This view holds that routine early cognitive development and the process of scientific discovery both result in knowledge of 'theories;' in particular, it is claimed that the child's 'theory of mind' really *is* a theory. I will discuss this version in the next section where I conclude that it is not useful to insist that things which are merely theory-*like* really *are* theories.

The third version of theory-theory goes deeper than the first two because it tries to account for the nature and acquisition of concepts. In its most explicit and sophisticated form, developed by Carey (1985, 1988), fundamental structures of thought are said to depend upon 'ontological' concepts, such as PHYSICAL OBJECT, LIVING THING, and so forth. The identity of an ontological concept is determined by the role it plays in a set of explanatory principles grasped by the child. A given set of explanatory principles is domain-specific and theory-like, but, most importantly, constitutes the 'packet of knowledge' that allows the child (or other user) to pick out just those things in the world to which the concept refers. Put more simply, a concept, e.g., DOG, is possessed by grasping a certain commonsense theory, namely, the theory that tells the user what kind of thing a *dog* is. Acquiring this concept is acquiring the theory of what a *dog* is. If (knowledge of) a given theory is innate, then the associated concept will also be innate; if a given theory must be acquired, then the associated concept must be acquired (by acquiring knowledge of the theory). Perner (1991, 1995) has applied this framework to the concept, BELIEF. In his account, the child acquires the concept BELIEF by acquiring a theory of what *beliefs* are, namely, the theory that *beliefs are representations*. I discuss this version of theory-theory in a later section, pointing out that it requires the child to have obscure knowledge for which there is no independent evidence and that it still fails to account for possession of the concept, BELIEF.

Some current beliefs about BELIEF

The empirical basis of the belief problem is as follows. Wimmer and Perner (1983) developed a test of false belief understanding (the Maxi task) which showed that the majority of six-year-old children could pass, while four-year-olds performed at chance. Baron-Cohen, Leslie and Frith (1985) subsequently modified this task, simplifying it (the Sally and Ann task, Figure 1). They found that the majority of normally developing four-year-old children passed this version. This study also found that a majority of mildly retarded children with Down's syndrome could pass the task, but that children with autism, even with normal IQ's, failed. Subsequently, numerous studies have confirmed and extended these results (for reviews, see Happé, 1995 and Leslie, in press). By age four, most normally developing children are demonstrably employing the concept BELIEF.

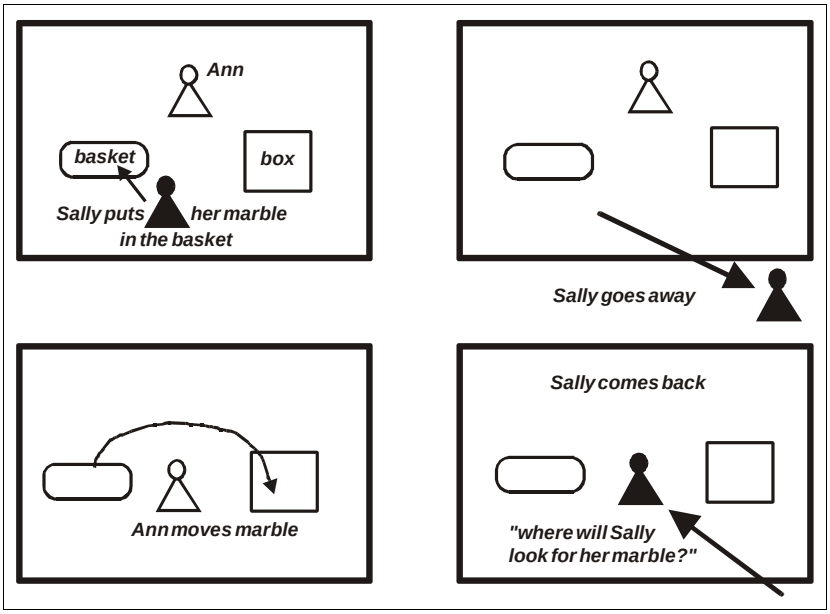

Figure 1 A standard test of false belief attribution. In addition to the prediction question shown here, children are asked two control questions, one to check that they remember where Sally put the marble and one to check they realize where the marble is currently. An alternative to the prediction question is the *think* question: Where does Sally think the marble is? Both prediction and think questions yield similar results with normally developing children and with children with a developmental disorder. (After Baron-Cohen, Leslie, & Frith, 1985).

The real-theory-theory

One version of theory-theory is that people, including children, 'have theories.' As I indicated, there is little in this claim to disagree with, in part because the notion of 'theory,' especially when extended from science to commonsense, is vague enough to cover almost any kind of knowledge and opinion.

Recently, however, the claim has been pushed to an extreme in which routine cognitive development and the process of scientific discovery are claimed to be essentially identical (e.g., Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1995).[3] Although there is some disagreement within the theory-theory camp over whether the child-as-scientist claim relates to the *process* of development (denied by Wellman, 1990 and by Perner, 1995 but espoused by Gopnik and Wellman, 1995 and by Gopnik and Meltzoff, 1997) or only to the *outcome* of that process (Perner, 1995; Wellman, 1990), there appears to be agreement that it relates at least to the outcome. Gopnik and Wellman (1994, 1995) develop their claim by thinking of scientific theories as a species of psychological entity. They are not concerned with the substance of any particular scientific theory, but rather with the general psychological properties of that whole class of knowledge. From this point of view, they generate a list of critical properties. The critical properties of scientific theories are said to be abstractness, coherence, predictiveness, defeasibility, interpretation of evidence, and explanatoriness. Gopnik and Wellman then point to features of the child's 'theory of mind' as it develops from about two to four years of age that illustrate each of these properties. They conclude that therefore what the child has acquired over this time really *is* a theory because these properties of a scientist's knowledge are also properties of a child's 'theory of mind' knowledge.

The case of language

Unfortunately, the properties that Gopnik and Wellman (also Gopnik and Meltzoff, 1997) consider crucial to establishing their claim fail to distinguish knowledge entities that are indisputably real theories from knowledge entities that are merely 'theory-like.' Consider the case of language. The left-hand panel of Figure 2 mentions an indisputably *real* theory of language, namely, the *Principles and Parameters* theory of generative linguistics (e.g., Chomsky and Lasnik, 1995). This theory is widely regarded as being a piece of *echt* science even by those who do not regard it as being true. Furthermore, it is undoubtedly the case that some people (certainly not me) possess real knowledge of this theory. So here is a clear sense in which someone (e.g., Noam Chomsky) knows something and the something that he knows really *is* a theory.

The right-hand panel of Figure 2, by contrast, shows the psychological entities and mechanisms that (are postulated by the theory on the left to) embody the knowledge of language that people routinely possess,

[3] For critical discussion of this idea see Carey and Spelke (1996), Leslie and German (1995), and Stich and Nichols (1998).

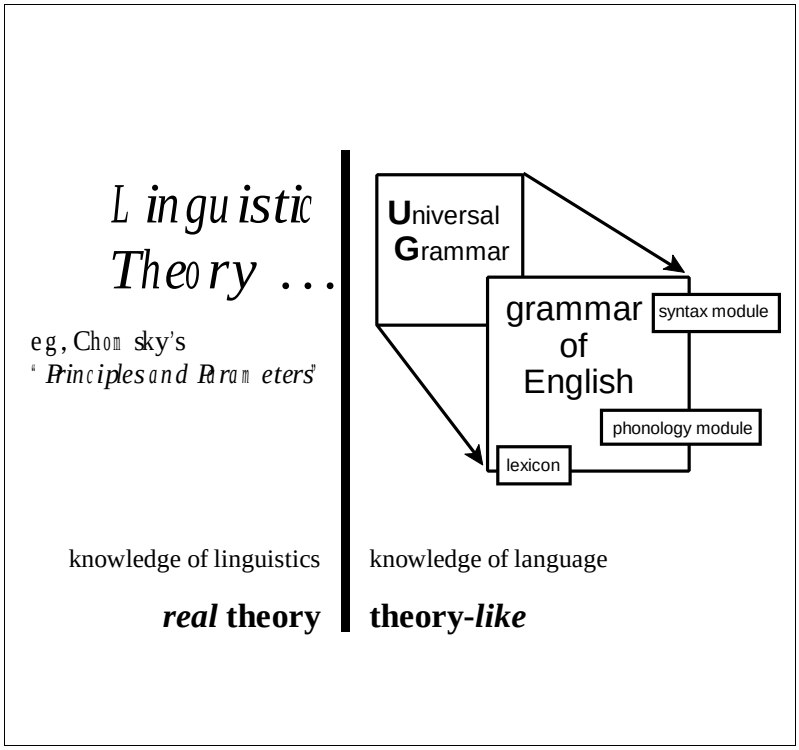

Figure 2 The case of language illustrates the distinction between a scientific theory ('real theory') and psychological entities that are theory-like. Both can be represented in the brains of people who possess the relevant knowledge: knowledge of linguistics and knowledge of language, respectively. However, most people only have knowledge of language.

including regular people like me and my neighbor's child, and not just special people like Chomsky. One of these entities is the "grammar of English," in some way represented in my brain and in the brain of my neighbor's child. Another entity is "Universal Grammar" which, according to the theory in the left-hand panel, is the entity, again in some way represented in the brain, that enabled me and my neighbor's child to acquire our knowledge of the "grammar of English." Chomsky's brain in some way represents all the entities depicted in Figure 2.

Now, a mental grammar has often been described as an internalization of a theory of a language, and the child's acquisition of a language has often been described as being like a process of theory formation, e.g., "[in acquiring knowledge of a language] the young child has succeeded in carrying out what from the formal point of view . . . seems to be a remarkable type of theory construction" (Chomsky, 1957:56). The entities or processes on the right of Figure 2 can reasonably be described as 'theory-*like*.' However, one would have to be *completely* blind to questions of mechanism to say that an internalized grammar, along with Chomsky's *Principles and Parameters*, really *is* a theory. Although almost nothing is known about the psychological basis of scientific knowledge, the best guess is that the child's knowledge of language is distinct from Chomsky's knowledge of linguistic theory in just about every respect that a psychologist might be interested in, including the mental representations involved, accessibility, penetrability, the timing, time course, and process of acquisition, and the underlying brain systems. Such distinctions are missed if we say that both knowledge of linguistics and knowledge of language really *are* theories.

As noted earlier, Gopnik and Wellman (1994, 1995 and also Gopnik & Meltzoff, 1997) argue that the child's 'theory of mind' really *is* a theory because it meets a set of criteria derived from a characterization of real theories. Unfortunately, these criteria also characterize the theory-like entities in the right panel of Figure 2 every bit as well as they characterize the *real* theory in the left panel. Theories postulate abstract entities that explain phenomena (Gopnik & Wellman, 1995:260): the child's internalized grammar is thought to 'postulate' abstract entities, e.g., categories like S and NP, properties of parse-tree geometry, and so forth, that explain sentence structure. Theories exhibit coherence in a system of laws or structures (Gopnik & Wellman, 1995:260): the child's internalized grammar is thought to be a system of interacting rules and representations that generate the structures of his or her language ('systematicity'). Theories make predictions "about a wide variety of evidence, including evidence that played no role in the theory's initial construction" Gopnik & Wellman, 1995:261): an internalized grammar allows the child to produce and comprehend novel sentences that "played no role in the [grammar's] initial construction" ('productivity'). Theories can be falsified by their predictions, yet may be resistant to counter-evidence, may spawn auxiliary hypotheses, etc. (Gopnik & Wellman, 1995:262-3): such phenomena in relation to the construction of an internalized grammar are much discussed in the language acquisition literature. Theories "produce interpretations of evidence, not simply descriptions . . . of evidence" (Gopnik & Wellman, 1995:262): internalized grammars produce interpretations of sound patterns in terms of

meaning via intermediate levels of structure including phonology, morphology and syntax, and not simply descriptions of the sounds themselves. Finally, although a "distinctive pattern of explanation, prediction and interpretation" such as we have noted above for grammar "is among the best indicators of a theoretical structure" (Gopnik & Wellman, 1995:262), it cannot distinguish a child's knowledge of language from Chomsky's knowledge of linguistic theory.

Modules and theory-theory

Gopnik and Wellman are not unaware that their criteria of 'theory-hood' are too weak to do much work. In contrasting their theory-theory view with the "innate module view" of the child's 'theory of mind', they note,

> ". . . many kinds of evidence that are commonly adduced to support [theory-theory] or [modularity], in fact, cannot discriminate between the two. . . the fact that the representations in question are abstract, and removed from the evidence of actual experience is compatible with either view." Gopnik & Wellman, 1994:282.

The failure to identify a *formal* basis for distinguishing between 'theory-like' knowledge structures (such as might be found in modular systems) and knowledge of 'real theories' should not be surprising. The philosophical project to develop a formal theory of what makes a set of beliefs into a scientific theory has long been abandoned as hopeless, as Gopnik and Wellman are aware. Many sets of 'beliefs,' even the 'beliefs' of perceptual systems, are abstract, coherent, predictive, explanatory, and offer interpretations that go beyond the evidence. There is no great harm in calling these systems 'theories' or 'theory-like.' But it is hard to see what the point might be in arguing that these systems 'really *are* theories' unless there's some definite way to distinguish them from systems which 'really *aren't* theories' but which are merely theory-like.

Gopnik and Wellman (1994, see also Gopnik and Meltzoff, 1997) advance one property of theories that they say discriminate theories from modules, namely, 'defeasibility.' The notion of defeasibility in the philosophy of science refers to the willingness of a theorist to regard a proposition or theory as 'negotiable' or revisable, for example, in the light of evidence. According to Gopnik and Wellman, this property of real theories is also a property of the commonsense theories that they attribute to children. Presumably, what they mean is simply that children's 'real theories' are revisable rather than that children always *believe* that their theories are revisable. In any case, according to these authors, modules are not similarly 'defeasible.' In fact, Gopnik and Wellman go so far as to label modules 'anti-developmental' (1994:283), apparently because they believe that knowledge in modules cannot be revised. They are careful to point out that it is not the issue of innateness that divides theory-theory from modularity theory. Indeed, they hold that theory-theory needs to postulate innate theories, including in particular, an innate 'theory of mind.' But these innate theories are not fixed for all time; they are 'defeasible' and are often quickly revised by the child.

However, even the property of 'defeasibility' does not discriminate between 'real theories' and 'theory-like' entities such as modules (see Stich & Nichols, 1998). It is hard to know why Gopnik and colleagues have come to believe that modules are fixed at birth, unrevisable, and 'anti-developmental.' None of the major modularity theorists posit such properties. Take the Chomskean modules of Figure 2 (right panel) as an example. The Universal Grammar module has the job of 'revising' itself in the light of the properties of the language(s) to which it is exposed. It does this by setting the values of a number of parameters. This in turn affects the nature of the grammar module that is constructed for a particular language. These modules learn and in the process 'revise' themselves and no doubt will have mechanisms to recover from error. My point is not that Chomsky's proposal is correct, just that in proposing modular processes Chomsky did not somehow overlook the fact that his modules were learning mechanisms. On the contrary, for Chomsky, that was the whole point. To take a rather different example of a module, consider Marr's (1982) 'Object Catalogue' whose job is to recognize 3-D objects from arbitrary viewing points. A module that performs this job has to learn the 3-D shapes of literally tens of thousands of everyday objects and no doubt makes the occasional error-plus-revision along the way. Again, my point is not that Marr's theory is right, just that in making his proposal, Marr, as an important modularity theorist, was quite happy that his module could perform a prodigious feat of learning. Or once again, consider the lexicon which modularity theorists, like Fodor (1983), often assume is a module. Given that the adult lexicon contains many tens of thousands of items (Levelt, 1999) and that infant lexicons contain none, the lexicon must learn on a grand scale, with the occasional recovery from error (Carey, 1978).

Innate theories and general learning

Gopnik and colleagues claim that modules are 'anti-developmental.' Perhaps they mean that the degree of defeasibility is too low, that 'theories' can be *radically* revised while modules can't. Wellman (1990) argues that the child's initial theory of belief is that "beliefs are copies of reality" but that this theory is soon revised to become the theory that "beliefs are representations of reality." Perhaps this is an example of radical revision of which modules are supposed incapable. The issues here are far from clear. However, it does seem odd that children should have an innate theory that almost immediately requires 'radical' revision and indeed that receives such revising within a year or two. If the necessary revisions to the innate theory become obvious to the average child between two and four years of age after applying his limited reasoning abilities to the morsel of idiosyncratic experience available in that time, why, with its vast experiential resources of biological time and whole populations, were these revisions not glaringly obvious to the processes of evolution or whatever Gopnik and colleagues assume bestowed the innate theory? Why doesn't Nature just bestow the revised 'theory' and be done with it? These are interesting questions, but, as Scholl and Leslie (1999b) point out, there is no reason to suppose that early 'theory of mind' involves 'radical revision' rather than plain learning. It is obvious why Nature should bestow a module that will contain more information at the end of its life than it does at the start. However, it is far from clear how the 'representational theory of belief' contains *more* information than the 'copy theory of belief,' rather than simply being a 'better' theory. And it is quite puzzling why Nature should bestow a *false* theory when she could have bestowed a true theory.

Perhaps what Gopnik and colleagues really want to say about theories versus modules is that theories are acquired by mechanisms of general learning whereas modules are mechanisms of specialized learning. Thus, someone acquiring knowledge of Chomsky's linguistic theories would have to employ mechanisms of general learning. Meanwhile, (according to Chomsky's theory) a child acquiring 'knowledge of language' employs specialized modular learning mechanisms. There are many interesting issues here that would take us too far afield to pursue. However, the evidence with regard to purely general mechanisms in 'theory of mind' development does not look good. Which general learning mechanisms might be involved? Presumably, exactly those that are used in scientific theory building. If that claim seems too strong, we can weaken it: if not those responsible for scientific creativity, then the mechanisms involved are those mechanisms involved at least in learning about scientific theories, or, at the very least, those involved in learning about 'science' at elementary levels of education. These mechanisms for 'real' science learning are highly sensitive to IQ, meaning that we find large differences between individuals in their ability to benefit from science education. Indeed, IQ tests were specifically designed to measure such differences in general or 'academic' intellectual ability (Anderson,1992). Mildly retarded individuals— for example, those with IQ's around 64 — have an extremely limited ability to acquire even elementary scientific ideas. Yet, mildly retarded non-autistic individuals can pass standard false belief tasks (e.g., Baron-Cohen et al., 1985; Happé, 1995). It has been clear for some time, then, that 'theory of mind' development is substantially independent of intellectual level and therefore cannot depend solely upon general purpose learning mechanisms. More recent evidence, some of it from unexpected sources, has also supported the modular nature of 'theory of mind' (Langdon & Coltheart, 1999; Leslie, in press; Varley & Siegal, in press).

Before I leave the question, I want to remark upon one property that real theories always have. It is impossible to imagine a scientific theory that is not explicitly articulated in a natural or a formal language. For example, Chomsky's knowledge of *Principles and Parameters* theory is explicitly articulated in a number of books and articles. Anyone who claims knowledge of Chomsky's theory must also be able to explicitly formulate its propositions, and to the extent he or she cannot do this, we deny them that knowledge. Translating this property into the 'real theory-theory' framework, we should say that knowledge cannot really *be* a theory unless it is explicitly articulated in a declarative representation. This places a strong requirement upon knowledge that is to count as a 'real theory:' it demands that the child be able to articulate, for example, his theory of belief. Is this too strong a requirement to place upon knowledge of a theory? It is if we want to allow 'implicit' knowledge of theories. Now, I am all in favor of implicit knowledge in theory-*like* entities and of leaving open to empirical investigation the question of which properties of a psychological entity are theory-like and which are not. That's the point of using *metaphors*. But can Gopnik and colleagues claim that a psychological entity really, non-metaphorically, *is* a theory and then get to pick and choose the properties in respect of which this is alleged to be true? Although I don't think they can, I shall put aside my misgivings. I shall not insist that the child be able to state (even) his 'real' theories.

However, I *will* insist that the theory-theorist be able

to articulate the child's theory — as it were, on the child's behalf. The articulable content of the child's theory forms the central substance of the claim made by the theory-theorist. In the case of Gopnik and colleagues, it is hard to discern exactly what the child's theory of belief is: What is it the child 'thinks' when the child entertains his 'representational theory of belief?' Surely, the child's theory can't simply be, "beliefs are representations." Why would *that* really *be* a theory? Both Gopnik and Wellman focus on what the younger child does *not* understand, but say little to specify what the older child's view actually is. Among the theory-theorists, only Perner has addressed this important point. I discuss Perner's specific proposals in the next section, after I have outlined the third and most interesting strand of current theory-theory. This version uses a theory analogy to provide an account of the semantics of abstract concepts.

Concept as theory

From this point on in the discussion, we will no longer worry about whether a 'theory' the child might have really *is* a theory. We will be content merely if a piece of knowledge is theory-like. In this section, we will be concerned principally with Perner's proposal, and Perner is not, as far as I know, committed to the 'theory of mind' really *being* a theory, in the sense of Gopnik and her colleagues. Perner (1991, 1995) is, however, committed to the child acquiring an explicit understanding of belief-as-representation, to the notion of conceptual change, and to the idea that "each particular mental concept gets its meaning not in isolation but only as an element within *an explanatory network of concepts*, that is, a theory" (Perner, 1991:109), and, therefore, to the idea of concept-as-theory.

The basic idea behind concept-as-theory is as follows. With something as abstract as *belief*, the only way that you could think thoughts about *belief*s is if you have a theory of what beliefs really are. Beliefs don't look like anything, they don't sound like anything, and they are not found in some specifiable location, and so forth, so how are you (your cognitive system/brain) going to describe (to yourself/itself) what a belief is? An attractive answer is that you will need something theory-like to specify what a belief is. The theory has to be accurate enough in its description of what a belief is to ensure that the concept, BELIEF, which is embedded in the theory, does in fact refer to beliefs and not to something else. The description is what will determine what is picked out by the concept. So, if the description does a very bad job (of describing what a belief is), and instead describes, say, a desire or a toothache, then the associated concept will not in fact be a concept of *belief* but a concept of *desire* or *toothache*, as the case may be. So the exact nature of the associated theory is vitally important because this is what determines both the *sense* of the concept and what its *referent* will be.

Moreover, on the concept-as-theory account, acquiring the concept, BELIEF, is acquiring the theory that says what kind of thing *belief* is. If the child has not acquired the theory, then he will not be in possession of the concept; if he acquires a theory that so badly describes *belief* that it instead describes *desire*, then the child will have acquired the concept DESIRE instead. It makes sense, then, on this version of theory-theory to pay a lot of attention to exactly what the child knows about *belief*. Because what he knows or doesn't know about *belief*, will determine what concept he has. To put it round the other way, you can discover what concept the child has by discovering what he knows or doesn't know about *belief*. But before you can decide whether the state of the child's knowledge means that he possesses the concept BELIEF, you must first decide what the critical knowledge is. This means you must decide what are *the* critical features of the adult concept BELIEF — what it is we big guys know about *belief* that makes our concept pick out just the things that are *belief*s. If you are a theory-theorist, this critical adult knowledge must be our commonsense theory of what *belief*s are. From the adult theory of *belief*, the developmental researcher derives a set of criteria that will be applied to the child's knowledge. If the child meets these criteria, he must possess the concept; if he does not, he must lack the concept. Hence the theory-theorist's interest in setting knowledge criteria for concept possession (Perner, 1991: Chapter 5).

The concept dictionary model

As I noted earlier, abstract concepts are widely supposed to be abbreviations for packets of knowledge. The concept-as-theory is one variant on this view. Imagine our repertoire of concepts as a dictionary — a long list of items, each made up of two parts: a concept on the left and an associated theory/definition on the right. Almost all the variance in theories of concepts has to do with the nature of the entries postulated for the right-hand side of the list: necessary and sufficient conditions (definitions), a stochastic function over features (prototypes), rules of inference, or theories. In every case, however, the entry on the right functions as some kind of a *description* of

whatever the concept on the left denotes. Hence the term Descriptivism for this general view of concepts. A dictionary model might be held explicitly, in the sense that its entries are assumed to be mental symbols or implicitly, in the sense that the entries are assumed to be merely emergent properties. Either way, *possessing* a given concept means having the correct entry for that concept in one's mental dictionary; *using* that concept (as the meaning of a word or as an element of a thought) is gaining access to the associated entry; and *acquiring* that concept means acquiring that entry.

Just as a real dictionary provides characterizations of words in terms of other words, so in the dictionary model of concepts it is assumed that the items on *both* the left and right sides of an entry are concepts. A concept is given a definition (or a prototype, theory, . . .) that itself is composed of concepts. For example, the entry for the concept DOG might give the definition, DOG = CANINE ANIMAL. In a prototype theory, DOG will be characterized as a stochastic function over properties such as HAIRY, FOUR LEGS, SLAVERS, BARKS, etc. A theory-theory might show an entry that makes critical reference to a dog being a LIVING THING. In all these cases, the descriptive entries are assumed to be made up of other concepts, such as CANINE, ANIMAL, HAIRY, LIVING THING, and so on, each of which will have its own entry with an associated description in the dictionary. That the descriptive entry is formed by other concepts is an especially natural assumption for the theory-theory, because it is hard to imagine how a theory could ever be stated without using concepts. In all dictionary model accounts, but in 'theory-theory' accounts in particular, possessing, using, and acquiring one concept depends upon possessing, using, and acquiring other concepts.

The dictionary model has a number of attractive features but it has one major drawback. The everyday word dictionary depends upon the fact that its user already knows the meanings of most of the words in the dictionary. If this wasn't true, the practice of defining one word in terms of a lot of other words would get nowhere. A dictionary in an utterly foreign tongue offers no point of entry or exit. If we know none of them, we can never escape from the maze of words and the dictionary is useless. The same point applies to the dictionary model of concepts. If we come to know what a given concept is by learning its (theoretical . . .) definition, which is given in terms of a lot of other concepts, then we will need already to possess those other concepts and already be able to pick out the things in the world to which they refer. But those other concepts are known by way of *their* entries in the concept dictionary which are comprised of a lot of still other concepts, and . . . Because this cannot literally go on forever, there must be some concepts which are known, not by a defining entry in the dictionary, but by some other route. These are usually called the *primitive* concepts. Primitive concepts provide the floor or ground upon which all other concepts are ultimately defined. A primitive concept is not acquired by learning a description; otherwise we are back in the maze. But, if there is a way to acquire a concept *without* learning a description, then the whole dictionary model is called into question. For this reason, dictionary models assume that primitive concepts are unlearned, i.e., innate.

With a highly abstract concept like BELIEF, the dictionary model creates a dilemma for theory-theory. Either BELIEF is primitive and innate, or it is acquired. If it is innate, then either the concept is constituted by an associated theory or it is not. If BELIEF *is* established by an associated theory (and *is* innate), then knowledge of that theory too *must* be innate. If it is *not* so constituted, then BELIEF is an abstract concept that falls outside the scope of theory-theory. And now we should ask for which other 'theory of mind' concepts theory-theory is irrelevant.

Alternatively, if BELIEF is acquired, then we have to ask: What are the *other* concepts, the ones in the associated description/theory/dictionary entry that the child has to acquire in order to possess BELIEF? Once we have an answer to that, we will be obliged to ask the same question about each of *those* concepts: What are their associated theories? What are the concepts in *those* theories? We must press our inquiries until, finally, we get answers that contain only primitive concepts. When we reach the innate primitive concepts, each of those concepts will either fall outside the scope of theory-theory or be constituted by an associated innate theory.

We can now understand the dilemma that BELIEF creates for theory-theory. When we pursue our repeated rounds of asking which concepts make up the associated theory that establishes BELIEF, the answers can go in one of two directions. Either the concepts in the associated entries become *less* abstract than BELIEF, or they become *more* abstract. If we assume they should be less abstract, we will end up characterizing BELIEF in behavioral terms. Theory-theorists correctly want to account for the *mentalistic* character of 'theory of mind' concepts but cannot do this by claiming that children are behaviorists. Alternatively, if we assume that the concepts in the associated entry for BELIEF are more abstract than BELIEF, we will find that our account ends up chasing larger and larger numbers of more and more abstract concepts, most of them quite obscure, while the possibility of accounting for their acquisition slips further and further from our grasp.

A close up of the representational theory-theory

Perner (1988, 1991) proposed that the four-year-old child comes to pass false belief tasks by discovering the representational theory of mind and in particular the representational theory of belief. Younger children adhere to a different theory, namely, that people are 'mentally connected' to *situations*, a theory which is meant to preclude conceptualizing belief such that the content of a belief can be false. Older children then make a theoretical advance, discovering that beliefs are really representations; this advance creates a new concept, namely, BELIEF, and ushers in success on false belief tasks.

When Perner originally proposed the representation theory-theory, the idea was that the child discovered that mental states were like *other* representations — like pictures or models, for example. Perner wrote,

> "If we define representation . . . as I have done, then we use the word "representation" to refer to the representational medium (more precisely the state of the medium). For instance, in the case of a picture it is the picture (medium) that is the representation and not the scene depicted on it (content)" (1991:280).

The key development in the child's 'theory of mind' was then said to occur around four years when the child acquired the (supposedly adult-like and commonsense) theory that mental states are internal representations. This, in turn, was said to be achieved by the child coming to "model models" by "work[ing] out the notion that something (referent) is apprehended (represented) as something (sense)." (Perner, 1991:284).

As Leslie and Thaiss (1992) point out, the most natural supposition for a representational theory of mind is that children acquire a representational theory of belief by hypothesizing that beliefs are internal mental pictures. Sally puts the marble in her basket and makes a mental picture or takes a mental photograph of the marble in the basket. Then she goes away with her mental picture. While she is away, naughty Ann discovers the marble and moves it from the basket to the box. Now, Sally is coming back! Where will she look for her marble? Answer: Sally will consult her mental picture which will show her that the marble is in the basket. This idea is highly attractive for a number of reasons. First, it provides a series of thoughts that preschool children might actually have, avoiding obscure and ultra-abstract concepts. Secondly, it would explain how preschoolers come to have the concept BELIEF by learning about things, like pictures, that are visible, concrete objects rather than invisible 'theoretical' constructs. Thirdly, mother can show you pictures, she can point to them, count them, discuss and compare them with you; in short, she can tutor you about pictures in ways she cannot tutor you about beliefs. Finally, almost every picture or photograph you have ever seen is 'false' or out-of-date, making them ideal for learning about their representational proprieties — about how something (you, a *big* boy or girl) is represented as something else (a baby).

Coming to solve the false belief task by way of a picture theory implies that understanding an out-of-date picture is a sub-component of understanding an out-of-date belief. If a picture task is a component task, then it cannot possibly be harder than a false belief task and, if anything, ought to be easier. Using tasks adapted from Zaitchik (1990), Leslie and Thaiss (1992) showed that out-of-date pictures are not easier and, in fact, are slightly harder, at least for normally developing children. For children with autism, Leslie and Thaiss showed exactly the opposite is true (see also Charman & Baron-Cohen, 1992, 1995). Understanding out-of-date pictures is therefore neither a necessary nor a sufficient condition for passing a false belief task. These findings are a blow to the idea that the child "works out" that beliefs have a "representational medium" (for further discussion, see Leslie and Thaiss, 1992, Leslie and Roth, 1993, and Leslie, 1994).

In light of these sorts of findings, Perner (1995) abandoned his original version of representational theory-theory. Rather than having to master a *general* theory of representation, the child is now said to employ a theory of representation *specific* to understanding beliefs.[4]

---

[4] Slaughter (1998) claims that the dissociation between children's performance on false belief tasks and photographs tasks is predicted by Gopnik and Wellman's theory-theory on the grounds that "[a]lthough theory-building processes require general cognitive skills and resources, the resultant concepts, including mental representation, are held to be specific to the domain of folk psychology" (p330). It is hard to see what property of Gopnik and Wellman's views predicts that concepts/theories should be specific in this way. Certainly, the opposite is true of real theories which strive for as much generality as possible. Indeed, the representational theory of mind is exactly the attempt to treat mental states as instances of something more general, viz., as representations. Without this generality, it is not obvious even what is meant by 'representational' in the phrase 'representational theory of mind.'

(continued...)

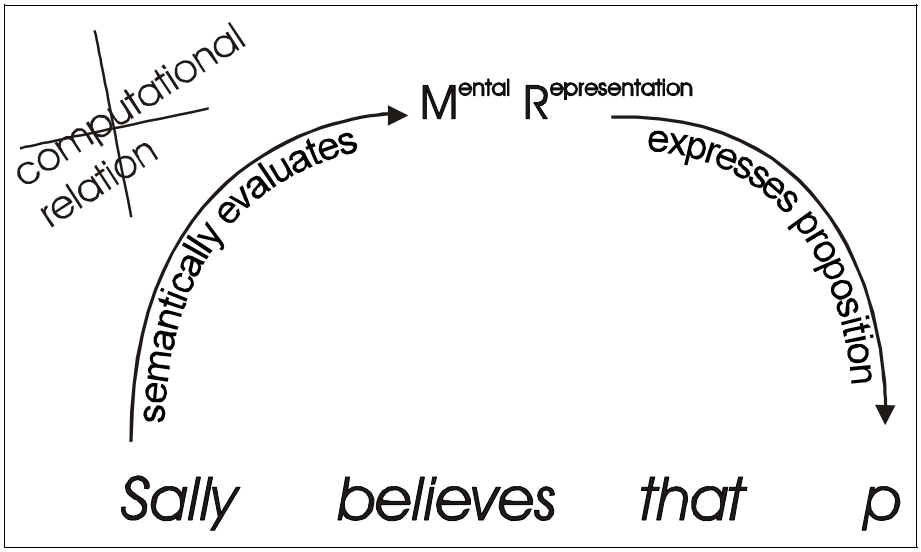

Figure 3: Perner's latest proposal borrows from cognitive science the idea that a belief is a relation to a mental representation. But instead of referring to a computational relation, the preschool child's BELIEF critical knowledge refers to a semantic evaluation relation to the mental representation. If Gopnik views the child as 'little scientist,' Perner views the child as 'little Fodor.' (After Perner, 1995.)

In characterizing the new theory-theory, Perner draws upon Fodor's explication of the theoretical foundations of cognitive science. Fodor (1976, 1981a) argues that a propositional attitude, such as *believing that p*, should be understood as a *computational relation* between an organism and a mental representation expressing the proposition *p*. Fodor's account is intended as a scientific account of what propositional attitudes really are. Perner attributes knowledge of this account to the child with one modification: instead of the child conceptualizing the notion COMPUTATIONAL RELATION, Perner says that the preschooler uses the concept SEMANTICALLY EVALUATES. According to Perner (1995), in order to understand that *Sally believes that p*, (in the case that *p* is false), the child must construct the 'metarepresentation,' *Sally semantically evaluates a mental representation expressing the proposition that p* (see Figure 3).

The first thing to notice is that moving from a general theory of representation to a specific theory of mental representation deprives the theory of any independent evidence. The 1991 version could hope to draw upon independent evidence that children first understand the idea of representing something as something else in regard to external, public representations like pictures, maps or models and then project these ideas to internal mental states. But, as we saw above, such independent evidence has evaporated. This has the disturbing consequence that the evidence supporting the idea that the child can understand that *Sally semantically evaluates a mental representation expressing the proposition that p* is just the evidence that supports the idea that the child can understand *Sally believes that p*, namely, passing false belief tasks. Therefore, there is, at present, no (independent) evidence to support the new theory-theory.

Let us remind ourselves of how Perner got to this position. He accepts the theory-theory account of concept possession: to possess the abstract concept BELIEF is to possess critical knowledge about *belief* and to acquire the concept is to acquire the critical knowledge. The critical knowledge in question is a theory of what *belief* is. In order to state the theory of *belief*, other concepts must be used. Therefore, the possessor of BELIEF must also possess these other concepts (the ones used to state the theory of *belief).* Rightly, Perner eschews the constraint that these other concepts must be *less* abstract than BELIEF. If, ultimately, BELIEF can be cashed out as or reduced to sensory concepts, then theory-theory is not really required. Moreover, reduction would entail that the child's (and our adult) 'theory of mind' concepts are fundamentally behavioristic and non-intentional. Rightly though, theory-theory is committed to mentalism. But rejecting this route, Perner is forced to allow the theory-

[4] (...continued)

Incidently, Slaughter (1998) overlooks the fact that in her study she compared children's performance on a modified photographs task with an unmodified false belief task. Just as it is possible to modify false belief tasks to make them easier for three-year-olds to pass, so it should be possible to modify photographs tasks too. Slaughter's results confirm this. According to the Leslie and Thaiss (1992) model, in making the comparison she did, Slaughter removed the only limiting factor that photograph tasks and false belief tasks have in common, namely, "selection processing." The resulting lack of correlation in children's performance does not "call into question the . . . model of development offered by Leslie" as Slaughter claims, but actually supports the model.

explicating concepts to be more abstract.

Perner is also forced to choose a theory of *belief* that might plausibly be true. The theory of *belief* that he requires has to explain how a thought containing the concept BELIEF actually picks out *belief* rather than something else, such as desire, serious facial expressions, an earnest gesture, or some other property of a situation containing a person with a belief. If the child (e.g., the three-year-old) has the wrong theory, then *his* concept BELIEF* will pick out something different from *our* concept BELIEF. And what theory can do the job of picking out *belief* other than *our* theory of what a *belief* really is?

However, there is a heavy price for taking this approach. In order to discover and apply the above representational theory of belief, the child must acquire the following concepts: SEMANTIC, EVALUATE, MENTAL, REPRESENTATION, EXPRESS, and PROPOSITION. The child *must* acquire these concepts because these are the concepts that state the critical theory of *belief*. Therefore, the child couldn't understand this theory unless he or she grasped these concepts. And if the child didn't understand this theory, then, according to Perner and theory-theory, the child wouldn't possess the concept BELIEVES.

We began by asking how one 'difficult' and obscure concept is acquired (BELIEVES), but now we have six more, each of which is just as 'difficult' and considerably more obscure. It is every bit as puzzling how the child might acquire any one of these six notions as it is puzzling how he acquires BELIEVES. One answer might be that these six concepts are innate. But if we are willing to accept *that*, why weren't we willing to accept that BELIEVES is innate? If we are not willing to accept these 'new' concepts as innate primitives, then each must, like BELIEVES, be acquired by acquiring and possessing critical knowledge — i.e., by acquiring a theory of *semantic evaluation*, a theory of *mental*, a theory of *representation*, and so on. Each of these theories will spin off further abstract concept-theory cycles, with no obvious end in sight. If we balk at *this* point in pursuing a theory-theory of concept possession and acquisition, the question inevitably arises why we didn't balk earlier at the first step: the decision to pursue a theory-theory of BELIEVES.

Unfortunately, the situation for the 'mental representation' theory-theory of *belief* is even worse than we have suggested so far. Fodor's formulation of propositional attitudes as computational relations to mental representations was designed to say what propositional attitudes *in general* are. It was not designed to characterize specifically *beliefs*. Fodor's formulation therefore does not distinguish *beliefs* from other mental states, such as *desires*, *hopes*, *pretends*, and so forth — they are *all* computational relations to mental representations. Each different attitude is assumed to involve a different kind of computational relation, putting it on the agenda of cognitive science to develop theories of each of the specific computational relations involved. This general characterization of propositional attitudes carries over into Perner's replacement of *computational relation* by a *semantic evaluation* relation (Figure 3). All propositional attitudes 'semantically evaluate' their 'mental representations'— by definition, propositional attitudes are attitudes to the truth of a proposition. So, even if the child *did* discover this obscure theory, it would still not provide him or her with the concept BELIEF, but only with an undifferentiated concept of *propositional attitude*. The theory in Figure 3 will only tell the child about propositional attitudes *in general*, applying to *desires* and *pretends* equally as it applies to *beliefs*. It will even apply just as well to *'prelief,'* the pretend-belief state that Perner, Baker and Hutton (1994) suggest three-year-olds attribute to other people. What it will *not* do is tell the child specifically what a *belief* is.[5]

Can the theory-theory in Figure 3 be patched up so that it provides to the child a theory of what *belief* is (as

[5] Fodor (pers. com.) points out that *believing that p* cannot be the same thing as *evaluating or holding-true a representation that means that p*. Consider: I have in my hands a copy of Einstein's paper on Special Relativity. I have never read this paper and, to be honest, I don't have a clue what it says. However, I know that the theory expressed in this paper is a cornerstone of modern physics, which, as far as I'm concerned, means that it's true. Secondly, this bunch of paper I have in my hands is only a representation of Einstein's theory. So I semantically evaluate (as true) this representation expressing the Special Relativity Theory. However, there is not a single proposition expressed in this paper that I have as a belief in the usual sense because I have no idea which propositions this paper expresses. But *whatever* they are, I hold them all to be true because I trust physicists to know what's what. But when I think that Sally believes that *p*, the marble is in the basket, I think that she actually grasps that very proposition. The idea behind treating belief as a particular kind of computational relation is that an organism standing in such a relation will *thereby* grasp and believe the proposition expressed. Without that assumption, a computational account of belief will not work. However, as the above example shows, exactly this assumption fails for the semantic evaluation relation. This is a further reason why substituting *semantic evaluation* for *computational relation* will not provide a theory of *belief*.

opposed to *desire*, *pretense*, etc.)? The main problem is with the relation 'semantically evaluates.' 'Mental representations that express propositions' will be a common feature in theories of *belief*, *desire*, *pretense*, *hopes*, etc. because Sally can desire that her marble be in the basket, can pretend that her marble is in the basket, or hope that her marble is in the basket, as well as believe that's where it is (and all the while the marble is in the box). What differs from case to case is the 'mode' of evaluation. The obvious temptation then is to add a simple qualification: Sally 'semantically evaluates *with respect to believing*' a mental representation . . . Certainly, this will do the job; but so will simply replacing 'semantically evaluates' with 'believes.' And both will work for exactly the same reason: namely, the concept BELIEF has been smuggled in. But this makes the theory-theory circular. *Belief* certainly is *belief* and, yes, the child will acquire the concept BELIEF by acquiring the concept BELIEF. But, theory-theories are simply not allowed to say that!

Agenda for a successful theory-theory

Here is a minimal agenda for a theory-theory of BELIEF. The first problem is to say, *without circularity*, what belief really is. Having made explicit the theory that constitutes the critical knowledge for concept possession, the next step is to provide *independent* evidence that the child does in fact acquire this critical knowledge. Finally, it must be shown that it is *by* acquiring this critical knowledge that the child acquires the target concept. Present accounts fall far short of achieving any of these goals.

Considerable obstacles lie in the way. I identified three sets of problems that face theory-theories of belief, and probably theory-theories more generally. The first is the *conceptual explosion* caused by concepts having 'dictionary entries' — that is, theories attached to each concept that have to say what sort of thing the referent of the concept really is, in order that the concept picks out that referent. Because theories are themselves composed of concepts, the number of concepts for which the theory-theorist must seek an acquisition account grows explosively. Proposing 'conceptual holism' sounds fine but I suspect that in practice it is quite frustrating. Secondly, because theory-theorists reject the idea that all abstract concepts reduce to statements formed solely of sensory concepts (you certainly can't be a theory-theorist if you accept that doctrine), the concept explosion will involve the *escalating obscurity* of the concepts that are spun off (cf. Fodor, 1981b). Perner's proposals illustrate this nicely: on the first iteration alone, we move from worrying about BELIEF to worrying about SEMANTIC. And *what* is the theory, grasped by the child, that constitutes the concept SEMANTIC? I suspect that escalating obscurity is a general feature of theory-theories — cf., DADDY = MALE REPRODUCER. Finally, the critical knowledge for BELIEF, conceptually rich and potent though it was in Perner's proposal, still fell short of the mark in specifying the exact meaning of BELIEF. This means that such a concept specification would not in fact do its critical job of specifically picking out *belief*s but instead would pick out any and all propositional attitudes. I suspect that the search for critical knowledge will only provide a *paraphrastic approximation* that forever falls short of its target — unless one introduces circularity.

Whether theory-theory can overcome these obstacles remains to be seen. In the meantime, alternative approaches should be vigorously explored.

Concept as soap molecule

One avenue to explore is dropping the notion that the *sense* of a concept — its associated critical knowledge — determines its reference and, therefore, which concept it is. The reference of a concept must be determined some how but stored knowledge is not the only conceivable way. In fact, it is far from clear how sense is supposed to determine reference. How does a concept fit to the world? Answer: A concept points to a stored description (of some sort); the description is laid against the world by the cognitive system and the things that fit the description are admitted to membership of the set of things in the concept category. But if a description is itself composed of concepts, saying that a concept fits to the world via a description does not answer the question of how a concept fits to the world. It provides a postponement instead of an answer.

Historically, there used to be a non-question begging answer. Empiricist philosophers, like Hume (1740), argued that concepts fall into two major types: the sensory and the abstract (in his terminology, "impressions" and "ideas," respectively). Possession of the sensory concepts is provided directly by the structure of the sensory apparatus. That is, Hume assumed that the way a concept like RED locked to the world did not entail applying knowledge of a description of *redness* (whatever that would be). The locking was provided by a mechanism, namely, the mechanisms of color vision. Such mechanisms can be innate and provide an innate concept RED without us having to suppose that *therefore* some piece of knowledge or theory about *redness* is innate. An

infant doesn't need a theory of *redness* if she possesses something else, namely, the mechanisms of color vision. Possession of the sensory concepts does not require knowing a dictionary definition or theory because these concepts are locked to the appropriate property in the world by sensory mechanisms.

So long as all non-sensory (abstract) concepts reduce to descriptions composed entirely of sensory concepts, we have a general outline for abstract concepts of how their sense determines their reference. But without that reductionist assumption about abstract concepts, we lack a proposal for how sense could determine reference. Theory-theory rightly rejects the notion that all abstract concepts reduce to sensory descriptions. But as we saw, this raises a number of obstacles to understanding how certain abstract concepts can appear so early in life. These problems could be avoided if there was some way for an abstract concept to lock to the world, other than through applying critical knowledge. In the case of sensory concepts, such an alternative has seemed uncontroversial: a sensory concept is locked to target by a psychophysical mechanism. Can this idea be extended to abstract concepts too?

The idea that certain abstract concepts might be acquired by way of a mechanism that locks to a specific target property in the world is certainly a wild idea. But is it wild enough to be true? There is a philosophical tradition (that has received far less attention than the mainstream Descriptivist accounts to which theory-theory is heir) which has tried to develop causal theories of reference (e.g., Fodor, 1998; Kripke, 1972; Margolis, 1998; Putnam, 1975). The fundamental idea is that concepts bear information about a specific property, not because of subjective knowledge, but because of an entirely objective causal relation between the concept (as psychological entity) and a property ('in the world'). Such views stress the 'psycho-physical' duality of concepts. Like a soap molecule with one pole locked to oil and the other pole locked to water, a concept has one 'pole' locked to the causal processes of a cognitive system and the other 'pole' causally locked to the world. Instead of being lost in the endless maze of mutual inter-definition, the representational relation between concept and world is brought directly to the fore.

What is the role of knowledge in 'conceptual psychophysics?' Knowledge about the referent of a concept is acquired and associated with the concept, but this stored associated knowledge does not provide or constitute the sole locking mechanism for the concept. So the knowledge is free to change without affecting what the concept designates.

But isn't it true that we acquire new knowledge and that this new knowledge changes the way we think about something? Don't we learn about *beliefs* or *daddies* or *dogs* so that we come to see them in a 'new light?' Most certainly we do. What is at stake is not *whether* we learn, or whether that learning leads us to 'conceive' of things in a new way. What is at stake is whether, once our concept DADDY is locked to the world, its reference changes systematically in relation to our evolving ideas about what a *daddy* really is. According to the Descriptivist view (and theory-theory), it does. According to conceptual psychophysics, it does not. We can capture our strong intuition about changes in the way we 'conceive' of things by distinguishing between *concepts* and *conceptions*. 'Concept' will refer strictly to the symbol *cum* reference-relation-to-a-property, while 'conception' will refer to any knowledge associated with the symbol. 'Conception' will capture what we know or believe about whatever the concept refers to. Since what we believe about something determines how it appears to us, we can retain the intuition that new knowledge changes how we think about things. What new knowledge will *not* do is change the meaning of our concepts.

In theory-theory, or any Descriptivist approach, the claim that a given (abstract) concept is innate, entails that critical knowledge is innate. In a conceptual psychophysics framework, this entailment does not hold. A concept may be innate if at least one locking mechanism is innate (there does not have to be a unique or 'critical' mechanism). The existence of innate knowledge remains an empirical question, of course, and it is even possible that innate knowledge may play a role in a given locking mechanism. Likewise, in theory-theory, or any Descriptivist approach, the *acquisition* of a given concept entails the acquisition of critical knowledge. Again, this entailment does not hold within a conceptual psychophysics approach. Acquiring a new concept will mean acquiring a lock on a new property.

It seems to me that a good way to study these questions empirically is concept by abstract concept. Although there are a great many concepts, it would be a great advance to have an account for even a single abstract concept of how it is innate or how it is acquired. There have already been suggestive findings. For example, Leslie and Keeble (1987) showed that six-month-old infants recognized a specifically causal property of events in which one object launched another by colliding with it. They proposed that infant recognition was based upon a modular mechanism operating independently of general knowledge and reasoning to "provide information about the spatiotemporal and causal structure of appropriate events" and that "it could do this without having to know what a cause 'really' is" (p.286). Such a mechanism

would allow the infant to attend to physical causation, to lock in the concept CAUSE, and then begin to learn about causal mechanisms from instances. There are also promising ideas concerning locking mechanisms for number concepts (see, e.g., Gallistel & Gelman, 1992) and faces (Johnson & Morton, 1991). Recently, Leslie, Xu, Tremoulet & Scholl (1998; see also Scholl & Leslie, 1999a) have suggested an account of how the infant's concept of *object* gets locked without recourse to knowledge of a theory of objecthood. Finally, in the 'theory of mind' domain, Leslie (1987) proposed a model of how the concept PRETEND is locked without assuming that the infant has critical knowledge of what pretending really is (see also German & Leslie, unpublished). In a similar vein, Leslie (in press) discusses the development of the concept BELIEF as part of a mechanism of selective attention.

So, how do you acquire a representational theory of mind?

In the theory-theory account, the child discovers a theory of general (or alternatively, mental) representation that gives birth to the concept BELIEF and to success on false belief problems. In this chapter, I have laid out a number of reasons that make me skeptical of this claim. In fact, I think the relationship between concept and theory is exactly the reverse. It is the possession of the concept BELIEF (plus a gradual increase in skill at employing the concept) that eventually gives rise to a commonsense representational theory of mind. As the child begins to enjoy increasing success at solving false belief problems, he or she will increasingly *notice* false beliefs and the circumstances that give rise to them. In an everyday sense, the child will then develop commonsense 'theories' about how other people represent the world. For example, if Mary represents bananas as telephones, the child can model this fact as *Mary thinks bananas are telephones*. Or if the child sees a dog chase a squirrel which then runs up tree B, while the dog goes barking up tree A, the child can 'theorize' that *the dog is barking up the wrong tree because it thinks there is a squirrel up there*. In the limited manner of commonsense theory and opinion, this is a representational theory of the dog's mind. If it is disappointing to find that the child's 'representational theory of mind' is so mundane and epiphenomenal on the child's concept of *belief*, at least we know that children actually think thoughts like these. There is no evidence that children *ever* explicitly think thoughts of the sort in Figure 3.

References

Anderson, M. (1992). *Intelligence and development: A cognitive theory*. Oxford: Blackwell.

Baron-Cohen, S., Leslie, A.M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, **21**, 37–46.

Boyer, P. (1994). Cognitive constrains on cultural representations: Natural ontologies and religious ideas. In L.A. Hirschfeld and S.A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*. (pp. 391–411). Cambridge, UK: Cambridge University Press.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G.A. Miller (Eds.), *Linguistic theory and psychological reality*, Cambridge, Mass.: MIT Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (1988). Conceptual differences between children and adults. *Mind & Language*, **3**, 167–181.

Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, **63**, 515–533.

Charman, T., & Baron-Cohen, S. (1992). Understanding drawings and beliefs: A further test of the metarepresentation theory of autism (Research Note). *Journal of Child Psychology and Psychiatry*, **33**, 1105–1112.

Charman, T., & Baron-Cohen, S. (1995). Understanding photos, models, and beliefs: A test of the modularity thesis of theory of mind. *Cognitive Development*, **10**, 287-298.

Chomsky, N.A. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N.A., & Lasnik, H. (1995). The theory of principles and parameters. In N.A. Chomsky, *The Minimalist Program*. (pp. 13–127). Cambridge, MA: MIT Press.

Flavell, J.H., Green, F.L., & Flavell, E.R. (1993). Children's understanding of the stream of consciousness. *Child Development*, **64**, 387–398.

Fodor, J.A. (1976). *The language of thought*. Hassocks, Sussex: Harvester Press.

Fodor, J.A. (1981a). Propositional attitudes. In J.A. Fodor (Ed.), *Representations: Philosophical essays on the foundations of cognitive science*. (pp.177–203). Brighton: Harvester Press.

Fodor, J.A. (1981b). The present status of the innateness controversy. In J.A. Fodor (Ed.), *Representations: Philosophical essays on the foundations of cognitive science*. (pp. 257–316.) Cambridge, MA: MIT Press.

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

Fodor, J.A. (1998). *Concepts: Where cognitive science went wrong*. Oxford: Clarendon Press.

Gallistel, C.R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, **44**, 43–74.

German, T.P., & Leslie, A.M. (unpublished). Children's inferences from *knowing* to *pretending* and *thinking*. MS. Rutgers University Center for Cognitive Science.

Gopnik, A., & Meltzoff, A.N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H.M. (1994). The theory theory. In L. Hirschfeld and S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*, (pp. 257–293). New York: Cambridge University Press.

Gopnik, A., & Wellman, H.M. (1995). Why the child's theory of mind really *is* a theory. In M. Davies & T. Stone, (Eds.), *Folk psychology: The theory of mind debate*. (pp. 232–258). Oxford: Blackwell.

Happé, F.G. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development*, **66**, 843–855.

Hume, D. (1740). *A treatise of human nature*. London: Clarendon, 1978.

Johnson, M.H., & Morton, J. (1991). *Biology and cognitive development.* Oxford: Blackwell.

Keil, F.C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kripke, S.A. (1972). Naming and necessity. In D. Davidson and G. Harman (Eds.), *Semantics of natural language*, (pp. 253–355). Dordrecht: Reidel.

Langdon, R., & Coltheart, M. (1999). Mentalising, schizotypy, and schizophrenia. *Cognition*, **71**, 43–71.

Laurence, S., & Margolis, E. (in press). Concepts and cognitive science. In E. Margolis and S. Laurence (Eds.), *Concepts: Core Readings*. Cambridge, MA: MIT Press.

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

Leslie, A.M. (1994). *Pretending* and *believing*: Issues in the theory of **ToMM**. *Cognition*, **50**, 211–238. Reprinted in J. Mehler and S. Franck (Eds.), *COGNITION on cognition*, pp. 193–220. (1995). Cambridge, MA.: MIT Press.

Leslie, A.M. (in press). 'Theory of mind' as a mechanism of selective attention. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences*, 2nd Edition, Cambridge, MA: MIT Press.

Leslie, A.M., & German, T.P. (1995). Knowledge and ability in “theory of mind”: One-eyed overview of a debate. In M. Davies and T. Stone (Eds.), *Mental simulation: Philosophical and psychological essays*. pp. 123–150. Oxford: Blackwell.

Leslie, A.M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, **25**, 265–288.

Leslie, A.M., & Roth, D. (1993). What autism teaches us about metarepresentation. In S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen (Eds.), *Understanding other minds: Perspectives from autism*, (pp. 83–111). Oxford: Oxford University Press.

Leslie, A.M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, **43**, 225–251.

Leslie, A.M., Xu, F., Tremoulet, P., & Scholl, B. (1998). Indexing and the object concept: Developing ‘what’ and ‘where’ systems. *Trends in Cognitive Sciences*, **2**, 10–18.

Levelt, W.J.M. (1999). Models of word production. *Trends in Cognitive Sciences*, **3**, 223–232.

Margolis, E. (1998). How to acquire a concept. *Mind & Language*, **13**, 347–369.

Marr, D. (1982). *Vision*. San Francisco: W.H.Freeman & Co.

Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review*, **92**, 289–316.

Perner, J. (1988). Developing semantics for theories of mind: From propositional attitudes to mental representation. In J. Astington, P.L. Harris and D. Olson (Eds.), *Developing theories of mind*, (pp. 141–172). Cambridge: Cambridge University Press.

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

Perner, J. (1995). The many faces of belief: Reflections on Fodor’s and the child’s theory of mind. *Cognition*, **57**, 241–269.

Perner, J., & Wimmer, H. (1988). Misinformation and unexpected change: Testing the development of epistemic state attribution. *Psychological Research*, **50**, 191–197.

Putnam, H. (1975). The meaning of 'meaning'. In K. Gunderson (Ed.), *Language, mind, and knowledge*, (pp. 131–193). Minneapolis: University of Minnesota Press .

Scholl, B.J., & Leslie, A.M. (1999a). Explaining the infant’s object concept: Beyond the perception/cognition dichotomy. In (Eds.), E. Lepore & Z. Pylyshyn, *What is Cognitive Science?* (pp. 26–73). Oxford: Blackwell.

Scholl, B.J., & Leslie, A.M. (1999b). Modularity, development and ‘theory of mind’. *Mind & Language*, **14**, 131–153.

Slaughter, V. (1998). Children’s understanding of pictorial and mental representations. *Child Development*, **69**, 321–332.

Stich, S., & Nichols, S. (1998). Theory-theory to the max: A critical notice of Gopnik & Meltzoff's *Words, Thoughts, and Theories*. *Mind & Language*, **13**, 421–449.

Varley, R., & Siegal, M. (in press). A dissociation between grammar and cognition. *Nature*.

Vosniadou, S. (1994). Universal and culture-specific properties of children’s mental models of the earth. In L.A. Hirschfeld and S.A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*. (pp. 412–430). Cambridge, UK: Cambridge University Press.

Wellman, H.M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, **13**, 103–128.

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and 'false' photographs. *Cognition*, **35**, 41–68.