RuCCS TR-16
December, 1994

Knowledge and ability in "theory of mind": One-eyed overview of a debate

Alan M. Leslie
Rutgers Center for Cognitive Science and
Department of Psychology
Rutgers, The State University of New Jersey
aleslie@ruccs.rutgers.edu

Tim P. German
MRC Cognitive Development Unit
University of London
tim@cdu.ucl.ac.uk

Technical Report #16
Rutgers Center for Cognitive Science
Rutgers, The State University of New Jersey
Psychology Annex, Busch Campus
New Brunswick, NJ 08903

Knowledge and ability in "theory of mind": One-eyed overview of a debate

Alan M. Leslie

MRC Cognitive Development Unit
University of London
and
Center for Cognitive Science
Rutgers University

Tim P. German

MRC Cognitive Development Unit
University of London

Responding to Gordon (1986), Stich and Nichols (1992; hereafter referred to as S&N1) began a debate in the pages of *Mind & Language* between those who believe that commonsense psychology is simply an ability to "simulate" the behavior of other people and those who believe that our capacity to understand mental states is a kind of commonsense "theory". Our angle on this debate is to worry about the *capacity to acquire* a commonsense psychology or "theory of mind." We believe the capacity to acquire a "theory of mind" (ToM) is domain specific and innate. We will make no bones about the fact that we are on the side of theory-theory and that we are skeptical about at least *radical* simulationism. This then will be a one-eyed overview of the debate. We shall try to do two things. First, we shall characterize what we think the "big issue" between theory-theory and simulation is. Second, we shall show why simulation theory, if formulated so that it poses a radical challenge to theory-theory, is implausible and why simulation theory, if formulated more plausibly, though not without interest, is simply a version of theory-theory.

In their second article on this topic (hereafter S&N2), Stich and Nichols argue that the big issue separating theory-theorists and simulationists is the issue of what Pylyshyn (1984) calls *cognitive penetrability*. Put simply, a process is cognitively penetrable if knowledge or representation can influence the outcome of the process in a "rational" way, e.g., through entering into a sequence of inference. A process that cannot be so influenced is said to be cognitively impenetrable. The radical simulationist claim is that commonsense psychology is cognitively impenetrable to theory of mind knowledge because, in understanding the behavior of another person, one simply runs the action planning device that generates one's own behavior while "pretending" to be that other person. The device is run "off-line" without producing external behavior and one internally observes its pretend output. Having thus no need of ToM knowledge, simulation accounts claim that none exists.

In contrast to the above, a theory-theory must assume that at least some, presumably specialized, ToM knowledge both exists and influences at least some of the processes of understanding others' behavior. We complement S&N's discussion of these issues by focusing upon the capacity to acquire a "theory of mind." The big issue in this context is whether both knowledge *and* ability or simply ability alone is involved in acquiring a "theory of mind."

Knowledge and ability

Theory-theories of folk psychology—and SN1+2 are surely correct in pointing out that there can be many different versions—hold that ToM capacity comprises *both* knowledge *and* ability. The simulation view—and again there can be different versions—is distinguished by claiming that folk psychology comprises *only* ability. The (radical) simulation view makes a stronger claim than theory-theory in this regard, since theory-theory could include simulation as one of its associated abilities but not vice versa. Hence the principal strategy pursued by the radical simulationist is to argue that what appears to be ToM *knowledge* is actually just *ability*.

In a similar vein, it used to be argued that knowledge of language was really just a practical ability—a set of habits, a skill, or even "present dispositions to verbal behavior"—and that acquiring a language was just learning a repertoire of responses (e.g., a list of sentences). This approach to language proved sterile for reasons made explicit by Chomsky (e.g., Chomsky, 1959, 1965, 1975). The basic obstacle to this approach is that the language faculty forms a cognitive system that comprises *both* knowledge and ability. Language learning, for example, involves acquiring a structure of knowledge—a grammar—and not just a list of responses. Chomsky (1988) points out that knowledge of language and language ability cannot simply be equated. For example, language ability can improve with no gain in knowledge, e.g., already existing knowledge may be accessed more efficiently and expressed in a more polished performance. Conversely, ability can be impaired with no loss of knowledge. If Juan suffers a head injury and loses all ability to speak and understand Spanish, must he thereby have lost all knowledge of Spanish? Not necessarily: Juan may recover his ability after a few weeks without following again the acquisition process by which he first gained his knowledge.

In Chomsky's example, Juan retains a system of knowledge (i.e., knowledge of Spanish) while losing the ability to deploy it. It is to this system of knowledge that we appeal in explaining why Juan believes that *el libro* refers to a book and not to a table. Juan's not believing that *el libro* refers to a table is hardly a result of impaired ability or lack of skill on Juan's part. Rather it is due to a property of Juan's internal system of representation for Spanish that Juan believes *el libro* refers to a book rather than to a table. Moving to examples closer to our present concerns, Chomsky (1988:31) argues that the concepts labelled by words "do not constitute a mere list". Instead, "they enter into systematic structures based on certain elementary recurrent notions [such as, action, agent of action, goal, intent, etc.] and principles of combination". Chomsky points out some of the subtleties involved in understanding words such as *follow* and *chase*, where the latter but not the former necessarily involves intention on the part of the agent. Or words like *persuade*: to persuade

someone to do *x* is to cause them voluntarily to decide or intend to do *x*; to persuade someone that *p* is to cause them to believe that *p*. Knowledge of vocabulary is comprised, in part, of the representation of such systematic distinctions and combinations of elementary notions, while its acquisition is guided by pre-existing representations for the elementary notions, such as agent, action, goal and propositional attitude.

Whether or not ToM will turn out to be like language and involve systems of knowledge *and* ability is, of course, an empirical question. For our part, we expect an affirmative answer. For a start, so far as we know, all languages provide elaborate lexical and syntactic apparatus for expressing ToM-related distinctions. The child acquires this apparatus in parallel with the growth of his ToM-related knowledge and ability. Knowledge of language and the specific capacity to acquire knowledge of language can be identified with systems of internal linguistic representation. These systems make explicit information concerning the entities, relations, principles and facts of the language domain. Likewise, we can identify "knowledge of ToM" with the system of representation for the entities, relations, principles and facts of the ToM domain. The postulation of such a system of representation, including knowledge specifically required for the acquisition of ToM, is what will qualify an account as a "theory-theory" of ToM. The broad definition of "theory" found in SN1 is in agreement with this basic idea. Simulation theorists, if they adopt a sufficiently strong position, can hope to pose a radical challenge to theory-theory by denying the existence of any such knowledge. We think that such a strong version, to the extent it can be made clear, is implausible. On the other hand, we think that weaker forms—namely, those that allow knowledge as well as ability—are entirely plausible but are really just versions of theory-theory.

However, we also want to give early notice that we reject many of the assumptions made by theory-theorists in the developmental literature. These theory-theorists have generally failed to address fundamental problems in the acquisition of ToM knowledge and have simultaneously ignored the role of limited ability in early ToM performance. To set the stage for our discussion of both these misguided approaches, we briefly sketch in the next section our ideas on the Theory of Mind Mechanism (**ToMM**).

ToMM: *the specific innate basis of our capacity to acquire a theory of mind*

Together with colleagues, we have been developing a particular version of theory-theory which has the aim of accounting for the normal acquisition and growth of ToM knowledge and ability during the preschool years and also for the pattern of abnormal ToM development found in children with autism (e.g., Baron-Cohen, 1991; Baron-Cohen, Leslie & Frith, 1985, 1986; Frith, 1989; Frith, Morton & Leslie, 1991; Leslie, 1987b, 1988b; Leslie & Frith, 1988, 1990; Leslie, German & Happé, 1993; Leslie & Roth, 1993; Leslie & Thaiss, 1992; Roth & Leslie, 1991; for a short review, see Leslie, 1992; for a lengthier treatment of current ideas, see Leslie, in press *a, b & c*). Central to our version of theory-theory is the idea that the core of our capacity to acquire ToM knowledge is a system of representation we call the "metarepresentation" (Leslie, 1987).

The metarepresentation is a certain kind of data structure computed by our cognitive system. This data structure provides an "agent-centered" description of a situation. It achieves this by making

explicit four kinds of information: (1) it identifies an Agent [who holds] (2) an identified attitude [to the truth of] (3) an identified proposition [describing] (4) an identified aspect of reality. One of the earliest observable manifestations of the deployment of the metarepresentation is the normal human capacity for pretence which includes the capacity to understand the pretence of other people. The human capacity for pretence emerges between 18 and 24 months after birth. Thus, we can illustrate the metarepresentation by reference to the infant interpreting mother's behavior of talking to a banana by computing the following metarepresentation: **mother** PRETENDS (of) **the banana** (that) **"it is a telephone"**.

To fulfill its task, the metarepresentation must comprise a number of components. The first of these components specifies who the agent is. The second component specifies the (informational) relationship between the agent and the following two components: an aspect of reality coded by a "primary representation", and an imaginary situation coded by a "decoupled representation". A primary representation is simply a literal, transparent description of a situation that, for example, results from perception. In contrast, a decoupled representation is "opaque" in terms of the standard tests of existential generalization, substitution of identicals, and entailment of truth. These three aspects of opacity are reflected in the three fundamental forms of pretend play and respectively allow the counterfactual representation of imaginary objects, of object identity, and of object properties (see Leslie (1987) for a more detailed account of the *isomorphism* between opacity and the fundamental forms of pretence). Whatever properties of internal representation give rise to opacity phenomena and allow counterfactual reasoning, these are structural properties of the human mind by the second birthday. Primary and decoupled representations together with "informational relations" (attitude concepts) make up a more complex, relational structure. We refer to this structure with the term "metarepresentation". This machinery translates into a specific and limited understanding that allows the child, under certain performance limitations, to represent particular attitudes (for example, PRETENDS) that agents can take to (the truth of) information, and, again under certain performance limitations, to interpret behavior accordingly.

A specialized mechanism, which appears to be modular, and which we call **ToMM**, deploys the metarepresentation early in development (towards the end of infancy), when encyclopædic knowledge and general problem solving ability is still very limited. The early growth of **ToMM** has important consequences, among which is the ability to construe agents as entities which are sensitive to information (Leslie, 1987). As a result of biological pathology, a failure in the normal growth of this mechanism occurs in children who will later be diagnosed as autistic. This produces characteristic impairments in these children's social and communicative competence. This work is revealing some aspects of the relationship between knowledge and ability in the development of ToM and we shall return to the topic later. For the moment, we shall consider the claims of simulation theory.

Radical simulation

Although the strong version of simulation theory denies that folk psychology is anything more than ability, we are not entirely sure that, by the end of the debate, there is anyone still trying to defend the position, though we suspect that Gordon wants to do this and, perhaps at times, Goldman too. Harris, however, (at least on our reading of Harris, this volume), retreats from the radical position

and is willing to allow the representation of propositional attitudes (i.e. metarepresentation) to enter the scene fairly early in development (though not, we think, early enough). Harris's position then becomes a version of theory-theory with a mix of knowledge and ability, though he wants the mix to be mostly one ability and that one ability to be "simulation". However, Harris has in mind a notion of simulation that is very broad indeed, including almost any use of one's own knowledge in the interpretation of another person's behavior, including for example, using one's knowledge of English to understand what someone says to you. This will almost guarantee that most ToM abilities involve "simulation" but such an outcome is largely a terminological victory.

Terminology aside, theory-theories can easily accommodate such broad examples "simulation" abilities. Indeed, Leslie (1987) provided just such an example in his account of the early capacity to pretend and to understand pretence-in-others. Pretence emerges between 18 and 24 months of age in normal children and reflects an extremely early use of core ToM knowledge, characterized by the theory of the "metarepresentation". One key part of Leslie's (1987) account of early pretence postulated that infants used the "primary" knowledge they had acquired about the physical world to elaborate their own pretend scenarios and to understand the pretend scenarios communicated to them in the action, gesture and speech of other people. Previous writers had sometimes suggested that children had to "learn to pretend" by learning "pretend transformations" and by acquiring other specialized skills. For example, it was sometimes assumed that children would have to learn a "schema" for pretending to drink from an empty cup (they pretended was full), just as they had to learn a schema for dealing seriously with (really) full cups, or, at the least, they would have to learn to "transform" the latter schema into the former. Leslie's metarepresentational theory of pretence showed that this was unnecessary. Some simple, general assumptions about how processes of inference operate over the internal structure of metarepresentations shows how the child can employ his primary knowledge in pretend scenarios. For example, if the child can infer that a cup containing water will, if upturned over a table, disgorge its contents and make the table wet, then the same child can also elaborate his own pretence or follow another person's pretence using the same inference: if *x* pretends of the cup that "it contains water", and if *x* upturns the cup, then *x* pretends of the cup that "the water will come out of it" and "will make the table wet" (Leslie 1987: 418–419; see also further discussion in Leslie, 1988, in press*b*, and Leslie & Frith, 1990). These same assumptions (regarding the metarepresentation and inference) also account for the *productive* nature of early human pretending.

Now, if someone wants to call the above "simulation", then they can; but it adds little or nothing to the account to do so. On the other hand, you may ask, why call **ToMM** a "theory-theory"? The minimal answer is that, as we saw in the case of language, systems of representation themselves constitute bodies of knowledge. To fully deploy such systems, additional abilities are required (e.g., inferencing that is sensitive to the structure of the representations). For this entirely general reason, theory-theories embrace both knowledge and ability.

Theory-theories of ToM can accommodate trivially simulative abilities such as those discussed above; theory-theories can also accommodate more interestingly simulative abilities, such as those suggested by the experience of introspectively imagining how we would feel in someone else's shoes. However, it is far from clear that even this latter kind of simulative ability is entirely knowledge-free.

Likewise, SN1 (pp. 47–48) describe Fodor's view that the mechanism at the heart of simulation, namely, the action planning/decision system, has access to ToM knowledge. SN1 say they will treat this possibility as if it were a version of simulation theory. They admit that "it is a bit odd to draw the battle lines in this way", but remark that if they can still defeat simulation after this tactical manoeuvre, then so much the better for their account and so much the worse for simulation. However, we think that their tactical manoeuvre has an undesirable consequence. It makes the critical issue appear to be *where* in cognitive architecture ToM knowledge is located, rather than *whether* there is such a thing as ToM knowledge. If the action planning system is modular (as simulationists are presumably inclined to assume) but has access to a local encapsulated database or to a ToM-specialized representational system, then action planning itself will exemplify knowledge *and* ability. So SN1's tactic gives too much away. Fodor's suggestion is a Trojan horse with respect to radical simulation.

Less than radical simulation

Because claims about the "action planning system" play a central role in both radical and less-than-radical simulation theory, detailed and explicit accounts of this system are crucial. Unfortunately, such accounts do not yet exist. Current assumptions are probably too vague to support much analysis, but certain key problems can be brought into focus.

As we remarked earlier, Harris (this volume) adopts a less-than-radical simulation account. Whereas his position is compatible with theory-theory, we think those elements of simulation theory he does retain are unconvincing. We shall outline some of the difficulties they face. Harris apparently accepts a key role for metarepresentation in ToM development and therefore for ToM knowledge—e.g. he accepts that the child has access to concepts of propositional attitudes. However, he believes that metarepresentation somehow arises out of a more basic ability to simulate (or "pretend"), on the assumption that the more basic ability does not itself employ metarepresentation. Thus, in common with other simulationists (e.g., Gordon), Harris's view is that to understand that another person is acting with a given goal, you must "pretend" to be in that situation yourself and, by running your action planning system "off-line", to "pretend" to have that goal yourself.

Simulationists talk about "running the action planning system off-line" because the goal that results is not one you mean to act upon. However, "running off-line" is not quite as simple as it first appears. My action planning system surely comes up with goals that I do not act upon or do not mean to act upon, now or ever. Presumably, such goals are "off-line" too. However, these are still *my* goals (*my* off-line goals) as opposed to someone else's on-line goals that I simulate off-line. So something somewhere in the system has to carry the distinction between *my* goals (off-line or not) and someone else's goals. But because other people's goals can be off-line too, I need a way of distinguishing between someone else's on-line and someone else's off-line goals. According to simulation theory, even someone else's on-line goals have to be simulated by me off-line, so it's not clear how I simulate someone else's off-line goals (off-off-line?). In any case, at least two degrees of freedom are required and not just the one that simulationists customarily talk about. Keeping track of who has what kind of goal is one of many places where a representational system might come in handy. As we shall see presently, two degrees of freedom are not (nearly) enough.

So far we have considered the case of "pretending" to be someone else acting seriously and the case of "pretending" to be someone *considering* acting seriously—the off-off-line case. Now we have to add the case of understanding another person pretending, something which even young children manage to do. According to simulation theory, the only way you can understand someone else is to "pretend" yourself "into their shoes." But this raises obvious problems when what you want to understand is someone pretending, as Leslie (1990a) pointed out. When people pretend play, they sometimes act with pretend goals and they sometimes act with 'serious' goals in regard to pretend circumstances—for example, someone can *pretend* to upturn a cup that is really full of water, but someone can also *really* upturn an empty cup they pretend is full of water. We leave the reader to supply the other permutations. This means the action planning system has to simulate someone pretending to act with a serious goal as well as someone acting seriously in pursuit of a pretend goal. How does it mark these distinctions? The natural assumption to make is that the recursive properties of metarepresentations are exploited. However, this route is blocked either by the simulationist's adherence to the "ability-only" doctrine or, in Harris's case, by the need to derive meta-representations from a non-metarepresentational simulation ability.

If the inescapable recursiveness of mental state understanding is not to be explained by a representational system (because such a system is a system of knowledge), how is it to be explained? The only answer a simulationist can offer is in terms of a structure of ability-only knowledge impenetrable mechanisms. If my ability-only mechanism has to go off-line to handle my own pretend goals and also off-line to handle another person's serious goals, it will have to engage a different but embedded action planning system to handle another person's pretend goals (off-off-line goals). Even this will not suffice to distinguish, for example, my/your considering (own) goals off-line as part of serious decision making and my/your off-line goals as part of pretence, though these are not at all the same thing. Nor have we begun to say how beliefs and pretends (mine, yours) are distinguished by this system, but clearly it will need still more degrees of freedom than just mine/your/hers and on-line/off-line. Suddenly, the action planning system does not look so simple. Moreover, this extra machinery is required for doing theory of mind work, not for action planning—and yet we were supposed to get the theory of mind abilities for free!

But we have not finished. The simulationist's action planning mechanism will need a number of "modes" (one for each distinct attitude) together with a recursive functional architecture (with an embedded machine for each level of mental state content). This is because the functional architecture of the device must do all the work that a representational structure would "normally" do. There are over 200 attitude verbs in English, though there may be a few synonyms in there. As a rough guess, adults can easily handle about five levels of mental state embedding (e.g., it seems fairly easy for someone to follow the statement that John thought that Mary wanted Sally to persuade him that the hero of the film had hoped that his wife would not want to pursue her criminal career). Perhaps a singly embedded machine is just credible, but a doubly, triply, ... embedded machine is not. All this is simply to rediscover some of the things for which *representational* systems are eminently suited: variables, recursion, compositionality, and so on. Whereas we think that Harris is right to retreat from a radical simulation account, we think he has not retreated far enough.

Sometimes we feel that these issues are discussed by simulationists using a terminology that misleadingly creates the impression of offering a substantive alternative to existing theory-theories.

Thus, for example, Harris and Kavanaugh (in press) *say* they reject Leslie's metarepresentational account of early pretence but then make use of some of its key concepts under a different name. Indeed, Harris (in Harris & Kavanaugh) retreats yet further from radical simulation and does attribute structured representations to the young pretender, much as Leslie did. According to this version, pretence does not require "decoupled" representations but instead uses "flagged" representations. As far as we can tell, "flagged" representations have all the properties decoupled representations have except that there is no provision in the account for them to "belong" to anyone. Unfortunately, these free-floating "flagged" representations do not make much sense. Unlike decoupled representations, "flagged" representations do not form a component of a larger structure that represents an informational relation between an Agent and a "flagged" content. Yet an "informational relation"—i.e., a relation to the truth of the "flagged" content—is the only kind of relation that will do the work in this context. But apparently such relations are not represented in the "flagging" account. So the same free-floating "flagged" representations are used for representing other people's primary goals, other people's pretend goals, one's own pretend goals, and one's own primary (non-pretend) but not-to-be acted-upon goals, and so on—miraculously without anything else in the system keeping track of these distinctions!

Harris (1991) believes that a simulation process in early pretence will be *simpler* than a process that represents a propositional attitude. We think that one can hold such a belief only in so far as one is not required to spell out in detail just how the simulation process is to work in early pretence or if one ignores key phenomena. We do not think that we should deny infants access to propositional attitude representations because we have the feeling that such representations are somehow "too complex" for an infant's cognitive system. On the contrary, they provide an ingeniously straightforward solution to the difficult adaptive evolutionary problem of understanding the cognitive determinants of Agents' behavior.

Even if simulation processes can replace inferences, as sometimes they plausibly might, they still need essential *control* processes, with access to metarepresentations, to organize them and interpret their results. Goldman (1993) tries to find a way in which the action planning system could simulate recursively. It is not surprising, in light of the foregoing, that what he suggests makes extensive use of recursive *representations* (of propositional attitudes) for providing inputs to and representing intermediate products of the "simulation" process, as well as for interpreting its results. Thus,

> "...to simulate Mary [who believes that John believes that *p*, one will] generate some initial beliefs she would have about John. I put myself in Mary's shoes of agreeing with John that he will put away the chocolate. I feed an awareness of this agreement into my Mary simulation and allow an inferential process to operate on it. This inferential process outputs the conclusion that John will put the chocolate in some spot X and remember which spot it is. So I ascribe this belief to Mary..." (Goldman, 1993:107)

Apparently Goldman's "simulation" process uses inferences that operate over metarepresentations. This makes it a less-than-radical knowledge *and* ability account, where one of the abilities happens to be "simulation". Goldman concedes this, saying that he makes "no blanket

rejection of 'theoretical' inference in self- or other-ascription". Nevertheless, he suspects that simulation is where "the action is" or at any rate "most of it". Because, in our view, locating and quantifying "the action" will require detailed empirical investigation, arguing the issue in its absence is pointless. One thing, however, seems sufficiently clear already. Simulating mental states, in any interesting and plausible sense of the notion, requires the use of metarepresentation.

Less than credible simulation

We pointed out earlier that some definitions of "simulation" are so broad as to include almost any use of one's own knowledge. So construed, young children certainly "simulate": for example, they understand what a speaker says to them by accessing their own lexical representations (rather than consulting a representation of what the speaker's lexical representations are), though it adds nothing to existing accounts of language comprehension to call this "simulation". However, even if the term is used in this way, it is still the case that young children are by no means limited to "simulating". For example, Baldwin (in press) has recently investigated early word learning by ostension. Suppose an adult labels an object, say a chair, at a moment when the infant herself is looking intently at a cup. Does the infant think that the cup is called "chair"? Baldwin showed that around 18 months of age an infant will disengage her own attention from the cup and check on the focus of the speaker's attention. The infant then assumes that the word uttered refers to the object to which the speaker is attending. Presumably according to the simulation account, the infant has understood this by running her own action planning system "off-line". She 'pretends' that she herself had made the utterance while looking at the object, and, as a result of pretending this, is delivered of the notion that utterances made while looking at a given object refer to that object and therefore that the speaker *means* chair by saying "chair". As S&N2 point out, the action planning system must necessarily be an infallible simulator of itself—it is supposed to be the self same system when run normally and when run "off-line". However, neither children nor adults refer only to objects they are looking at. So if Baldwin's children use their own action planning systems to discover what the speaker means then they ought to know that people don't always refer to the objects they are looking at (for example, very often when people speak they look at each other!). It is far from clear how simulation provides an account of even this most elementary of ToM phenomena, computing speaker's meaning. Perhaps the infant assumes that if *she* were teaching someone the meaning of a new word then she would look at the object she named? But it seems hardly credible that the infant 'pretends' to be the speaker *teaching* the infant that speaker *means* chair by saying "chair"!

Any "theoretical" assumption the infant may makes is not at all guaranteed to be true. Though we rightly expect that the "theoretical" assumptions of commonsense to at least be useful, they are always potentially fallible. A much simpler theory-theory account of Baldwin's findings can be provided in terms of the infant employing a piece of fallible "theory". Now consider our infant further in terms of her capacity for pretence which emerges around the same time. This time her father playfully picks up a banana and speaks into it. The infant attends to this and smiles. Then the caregiver holds out the banana to the child and says, "The telephone is ringing. It's for you!". Fortunately, the infant does not learn from this that the word "telephone" can refer to bananas, despite the fact that father looks at the banana when he utters the word. Instead, the infant grasps the

fact that father is pretending that the banana is a telephone and interprets his speech accordingly. The infant calculates *speaker's* meaning in something like Grice's sense (Grice, 1957). Of course, that is what she did before in Baldwin's study, except there the speaker "really meant it" whereas now speaker only pretends to mean it. So this time, the infant has to "simulate" the speaker by 'pretending' to be someone *pretending* that "telephone" *means* banana. So many degrees of freedom to represent and, according to Harris, no system to represent it!

We expressed in the last section our reasons for skepticism about the existence of recursive ToM machines that operate without recursive representations. Now we can see that we should have to posit such systems in infants. We become yet more skeptical of this whole idea when Harris concedes that he wants to attribute recursive (propositional attitude) representations to children just a year or so older. We prefer our metarepresentational account, which maximizes continuity, to Harris's which maximizes change.

In summary, we have few qualms about entertaining the idea that "simulation" may be one of the ToM related abilities. What these abilities have in common is that they use structured, systematic metarepresentational knowledge. Access to metarepresentations is required to define the problems to be solved, to initiate and guide the problem solving process, to select relevant inputs for it, and to encode and interpret its intermediate and final results. This is consistent with the theory-theory view that commonsense psychology comprises both knowledge and ability. We see no reason to believe that simulation plays a fundamental *structural* role in ToM acquisition. On the contrary, simulation needs metarepresentation.[1] However, we should not be surprised if investigation showed

---

[1] There have been proposals recently that the ToM impairment discovered in the syndrome of childhood autism might reflect an impairment of "simulation". This suggestion has been made in two forms. The first is that autistic impairment is specific to simulation of the states of social agents (Harris, 1993). Presumably, on such an account, "simulation" should be required to understand the states of being happy and being sad. Yet autistic children seem to understand these states (Baron-Cohen, *et al.*, 1993). Presumably too "simulation" should be required to appreciate the distinction between moral and conventional injunctions. Autistic children make this distinction (Blair, unpublished). In fact, it looks as if the "simulations" autistic children have specific difficulty with are those that require metarepresentation. The second form the simulation-impairment-in-autism proposal takes is that "simulation" is a *general purpose* faculty and that autistic children are impaired in this general faculty (Currie, unpublished). This version of the account suffers all the difficulties of the first version plus some more. For example, use of visual imagery is apparently part of general simulation, yet autistic children perform normally on standard tests of visual imagery ability (Shah, 1988). Or again, when tested under the same conditions, autistic children can correctly calculate the content of an out-of-date photograph (drawing, map) but not the content of an out-of-date belief, (e.g., Chapman & Baron-Cohen, 1993; Leslie & Thaiss, 1992). Surely photograph and belief tasks both require "general purpose simulation (e.g., imagery)" if either does. Metarepresentational processes rather than "simulation" explains these patterns of impairments and spared abilities in

that "simulation" processes play other important roles, e.g. in moral persuasion, or in discovering through imagination what subtle emotional reactions one might have to a complex novel situation. If so, we shall still be in need of genuine theoretical insight into what "simulation" or "imagination" is supposed to be exactly. As regards radical simulation, we see no reason whatsoever to suppose that the psychology of the ToM domain is reducible to a ToM-knowledge-free ability. An engineer might use a pocket calculator in the course of building a bridge, but it would be a mistake to attempt to understand bridge building as nothing more than use of a pocket calculator. We think that the radical attempt to understand the ToM domain as nothing more than use of simulation is equally forlorn.

Problems with theory-theory

We turn now to consider some of the problems faced by current theory-theories. We think that part of the appeal that the "simulation" idea might have, for developmentalists at any rate, is the promise it makes of simplifying the knowledge that has to be attributed to the young child. Although we do not think it can deliver that promise in a radical fashion, we do think that ToM works well a lot of the time if you simply use your own knowledge about the world and that much of ToM development has to do with acquiring knowledge about when this does *not* work. If someone for some reason wants to call that "simulation", then we see little point in arguing.[2]

On the other hand, some theory-theories—encouraged perhaps by the phrase "*theory* of mind"—have claimed that the best way to understand preschool development in this domain is to view the child straightforwardly as a "little scientist". This has led to two sorts of claim: first, that the process by which the child develops ToM is very similar to or even the same process by which scientists develop their theories (e.g. Gopnik and Wellman, this volume); and second, that the outcome of this process, the knowledge acquired, is a sort of childish version of a scientific theory, in this case a particular scientific theory, namely, the Representational Theory of Mind (e.g. Perner, 1991).

The "child-as-scientist" metaphor raises a number of problems, some general in nature, some specific to this case. Among the general problems are the following: We have no clear idea what the process of scientific discovery is, and so can hardly use it to illuminate the process of development; good luck has at times played an important role in the unique history of science but can hardly enter as a factor in our account of cognitive development; the history of science has led many who have studied it to doubt whether there is a definable method for achieving scientific insight any more than

---

[1](...continued)
autism.

[2] We could always carry this a step further and argue that, because any mental content you attribute to someone else must necessarily be internally represented using one of your own mental structures, all attribution is necessarily "simulation". But again this seems pointless.

there is a definable method for achieving freedom and justice; science is largely a cultural-historical product depending upon unusual adult individuals painstakingly using all their culture's and all their personal intellectual resources, whereas the development of preschool commonsense is largely an early, rapid, and uniform expression of biological endowment. However, we put aside these and many other points that might be raised, because, despite our approximately equal ignorance of the processes of scientific discovery and of the processes of cognitive development, it may yet be possible to pursue particular parallels between cognitive development and the history of ideas as a way of suggesting problems that an explanatory theory of development should deal with (cf. Carey, 1988; Karmiloff-Smith, 1988).

But the moment one begins to use the metaphor of child-scientist as *explanation*, difficult and well known problems arise. For example, Gopnik and Wellman have the child-scientist "testing" his or her "hypotheses" about mind. If this is the case, the first thing we want to know is: Where do the hypotheses to be tested come from? Are there constraints on the class of admissible hypotheses or an ordering of admissibility? If there is not, why does the child in company with his peers "test" just the same specific range of hypotheses, given that, unconstrained, the class of hypotheses is infinitely large? Except to rule out "nativist" solutions (apparently on ideological grounds), these theory-theorists have little to say on these basic questions.

In addition to these general problems, there are specific difficulties faced by a version of the theory-theory that carries the child-scientist view to an extreme and attributes a "Representational Theory of Mind" (RTM) to the preschool child. The RTM account places paramount importance upon the finding that children do not pass certain tasks that require the correct calculation of content for false beliefs until they are aged four years or more. Failure before this age is interpreted as reflecting the lack of a concept of belief. This lack is in turn explained as a failure to understand that beliefs (and other mental states) are representations. The RTM account then shifts the problem from the acquisition of a concept of belief to the acquisition of a concept of representation. It is not clear what advantage is gained by this move because there is no account of where the concept *representation* comes from, and *representation* is hardly a less obscure concept than that of *belief*. However, an advantage might be gained if we assume that preschool children can get a purchase on the concept *representation* by somehow learning about public representational artifacts, like pictures, photographs, and maps (Zaitchik, 1990; but see Leslie & Thaiss, 1992). Having thus acquired the concept, *public representation*, the child must somehow gain the insight that mental states, and in particular, *beliefs*, are really *representations* too. No-one has any idea how the child might get such an insight, but if she did, she would be said to have constructed a RTM.[3]

---

[3] The foremost proponent of this view is Perner (1988, 1991) who has laid out at some length what he intends by attributing a Representational Theory of Mind to the preschooler. Perner's position rests upon the standard distinction between representations and propositions: representations are physical embodiments or expressions of propositions, the latter being, roughly, abstract "meanings". The three year old is supposed to have access to something like propositional attitude concepts (sometimes called a "situation theory" by Perner 1991), but Perner assumes that these notions are not powerful enough to allow an understanding of false beliefs. False belief

(continued...)

So the key notion on this story is that success on certain false belief tasks at 4 years is the result of a theory shift. Prior to the shift, the child (somehow) constructs a theory of representation by learning about artifacts like pictures (models, maps etc.) which, being both public and observable, are presumed to be easier to learn about than beliefs. Having developed a theory of "public" representation, the child applies it to the mind in the form of, for example, a pictures-in-the-head theory of mental states. A critical assumption, then, of this version of RTM—a version that at least has some prima facie plausibility—is that understanding public representations should occur earlier than understanding false belief.

This story can be given a little more sophistication: Although a photograph is a representation, it cannot be false in quite the way that a belief can be false. A photograph is always an accurate representation of *some* situation (e.g. the chocolate sitting in the cupboard). If the situation changes, the photograph is a still-accurate representation of that old situation, not a *mis*representation of the new situation. In the false belief task, Sally's belief starts, like the photograph, as an accurate representation of the situation. When the situation changes, however, unlike the photograph, Sally's belief *does* become a misrepresentation of the new situation. This is because Sally mistakenly *believes* that her representation (of the previous situation) accurately represents the current situation. The photograph cannot perform this trick because the photograph cannot *believe* anything.

Notice that the difference between the two cases above is precisely related to the special nature of *believing* (and more generally, to the special nature of propositional attitudes) rather than to the general problem of the nature of representations. The critical point is whether or not Sally *believes* her mental representation.[4] Understanding representation then could only be a

---

(...continued)
understanding only becomes possible later, at four years, as a result of a radical "theory shift" to a RTM. Leslie & Thaiss (1992) criticize these ideas. Although other writers in the RTM camp (e.g. Flavell, Gopnik, Wellman) align themselves with Perner, they fail to make clear whether they too are drawing a distinction between proposition and representation. It may be that, for these other writers, notions like description and representation can be used indiscriminately. But if all that is important about a "representation" for these writers is that, like a proposition, it is *semantically* evaluable, then it is hard to see in what the radical theory shift would consist. Everybody, as far as we know, agrees that even very young children view the intentional states and intentional behavior of agents as semantically evaluable. As we point out below, the key question for deciding whether to attribute a RTM to four-year-olds is whether or not they individuate beliefs formally or "syntactically", *as opposed to* semantically.

[4]It is easy to go round in circles trading on the ambiguity of the term "representation." For example, the term can be used widely so that beliefs and pictures are both examples of "representation", or it can be used narrowly so that it is synonymous with "belief". In the narrow sense, 'John represents the situation (to himself) as p' can be used to mean simply 'John believes p of the situation'. It is clear that if the RTM view of the child slips from the wide to the narrow sense it is certainly true but trivial. For now the RTM view says that children acquire (lack) a concept of

(continued...)

subcomponent of understanding belief. Unlike the photograph itself, which lacks the capacity to believe anything, Sally *could* mistakenly believe that the photograph depicts a current situation. On this account, the problem of understanding out-of-date representations (like old photographs or old pictures-in-the-head) is included as a subcomponent in the problem of understanding out-of-date beliefs. Thus, the concept of false belief includes all the conceptual complexities of representational pictures plus some other unspecified complexities specific to belief. Again, reinforced by the idea that public representational artifacts will be easier to learn about, this predicts that out-of-date pictures will be understood earlier than out-of-date beliefs.

Unfortunately for this account, the evidence from preschool development contradicts the prediction. When tested in the same way, out-of-date belief is understood *earlier* and not later than out-of-date pictures (Zaitchik, 1990; Leslie & Thaiss, 1992). Sally is replaced by a (Polaroid) camera and Sally's belief by a photograph taken by the camera of the marble in the basket. The photograph is then placed face down so that the child does not get to see it (after all, the child does not get to see Sally's belief). Now the marble is moved from the basket to the box and the child is asked where in the photograph is the marble. Most three-year-olds fail both out-of-date belief and photograph tasks and most four-year-olds pass both. But if a child passes only *one* of these tasks it is reliably false belief and not photographs (Leslie & Thaiss, 1992).

In light of the above, either (a) one must find an analysis in terms of general processes of theory construction in which out-of-date belief is *less* complex than out-of-date representation, or (b) one abandons assumptions in favor of purely general processes and instead looks for an account of belief understanding in terms of specialized, domain-specific mechanisms. If one opts for the first of these, one cannot account for the striking performance of autistic children which is near ceiling on out-of-date pictures but severely impaired on out-of-date beliefs (Leslie & Thaiss, 1992). These data are shown in Figure 1. This leaves the second option, a domain-specific mechanism, to which we return in the section after next.

---

[4](...continued)
belief because they acquire (lack) a concept of belief. The present discussion of RTM is directed toward the substantive wide construal and ignores the narrow vacuous equation for obvious reasons.

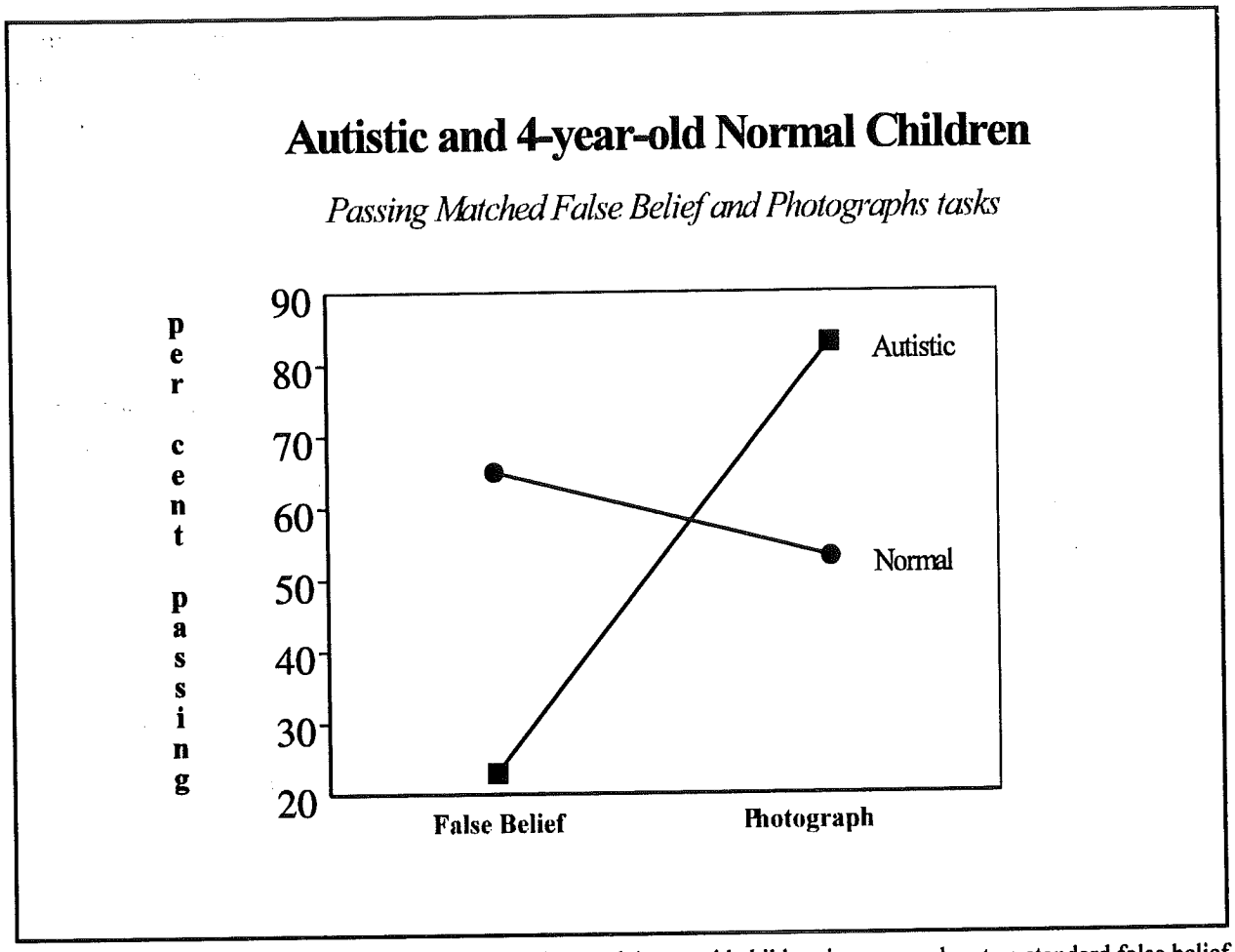

Figure 1. The performance of autistic children and normal 4-year-old children is compared on two standard false belief tasks and on two matched photographs tasks. The autistic children do significantly worse that the normal children on the false belief tasks and significantly better than the normal chldren on the two out-of-date photogrpahs tasks. After Leslie & Thaiss, 1992.

An insidious ambiguity

Although there is strong evidence against preschool RTM, it is nevertheless instructive to reflect that there has never actually been a shred of evidence in its favor. This view has gained the acceptance it has largely due to the ambiguities that surround the term "representation." We noted one of the more obvious ambiguities earlier (see *fnt.* ?). The ambiguity we draw attention to here is less obvious and, correspondingly, more insidious. The cardinal finding upon which arguments for the RTM view have been based is the four-year-old's new-found success on certain false belief tasks. Yet the falseness of a belief is a semantical property and semantical properties distinguish between beliefs as propositional attitudes, not between beliefs as representations.

Propositions are abstract 'meanings' and, because of this, they are distinguishable only on the basis of meaning. Two propositions are distinct, if and only if, they differ in *content*, regardless

of the form in which they are expressed. Representations, on the other hand, are physical embodiments of propositions—they are physical objects, states or events, like pictures, brain states, or utterances, that express propositions. Because of this, representations are distinguished one from another by their physical forms, even where they express the same proposition. Thus, I have two photographs, even though they are both photographs of the same scene on the same day. Likewise, two sentences in distinct languages are distinct as sentences (representations) even though they may have the same meaning—even though they express the same proposition. They will be distinct because they will differ in form, e.g. one photograph is larger than the other, the sentences differ in sounds when spoken or letters when written, etc. In short, propositions are individuated by content, representations by form.

The above distinctions carry over into two different approaches to mental states. Thus, in RTM, mental states are individuated in terms of their form—their "syntax"—rather than in terms of their content or "semantics". For example, in cognitive science, whence the idea of a RTM comes, the key questions concern the form or syntax of representations involved in mental states. We want to know the "format" of internal representations, e.g., whether something in memory is stored as a verbal expression or a motor program, whether there are visual images, whether they have an analogue or symbolic format, and so on. Thus, if John and Mary both believe that the same cat is sitting on the same mat, but John entertains this thought as a mental image of the cat on the mat, while Mary entertains an English sentence about the cat on the mat, then from the cognitive science (i.e., RTM) point of view, John and Mary have different mental states. From the point of view of *commonsense* (i.e., PA) psychology, however, if John and Mary both believe the same things about the same cat, then they have the same mental state, the same belief. This is because the PA view individuates mental states *semantically* by content, and not syntactically by form.

The claim then that the preschooler shifts his theory of mind from a PA based theory to a RTM is fundamentally the claim that the preschooler shifts from individuating mental states on grounds of content to individuating mental states on formal syntactic grounds, regardless of content. Recall what the basis of this claim is: the four-year-old's grasp of false belief. But whether a belief is true or false is a semantic question *par excellence*. Individuating beliefs and other mental states in terms of their content—does Mary believe the same thing as I do?—appeals to notions drawn from a PA based psychology. To demonstrate possession of a RTM, by contrast, we should need to show that four-year-olds will distinguish John and Mary's beliefs about the same cat on the same mat on representational grounds, that is, attribute *different* beliefs to John and Mary on the grounds that their mental states are formally distinct *despite the identity of their content*. To our knowledge, no study has ever demonstrated such a distinction in preschoolers. The only evidence that we know of speaks to the question of how preschoolers individuate mental states on semantic grounds. Thus, so far as we know, no evidence exists to support the hypothesis of a preschool RTM and no serious reason has ever been given for supposing that passing false belief tasks at four years has anything to do with a "radical shift" to RTM.

Perhaps, in light of the above, we will be told by the advocates of RTM-at-four that they did not, after all, mean *representation* and that, yes, RTM is a red herring. They might then add that the key shift at four years is to an explicit theory of the semantical notions of the *sense-reference* distinction and that, indeed, this is what the term *metarepresentation* "must" mean. But there is little

to gain from shifting the problem of the acquisition of the concept of *belief* to the problem of the acquisition of the twin concepts *sense* and *reference*. Again, these notions are hardly less obscure and certainly we have not been told where *these* notions might come from nor how they might be represented.

Leslie (1987) proposed the existence of an innate processing mechanism (the "decoupler") that operates from around 18 months and characterized it with regard to the phenomena of early pretence. This mechanism, together with the appropriate representational structures, *handles* the "sense-reference" distinction for the child, at least as far as it is relevant to early understanding of PA's (e.g., the opacity problem). All this was done without confusing the issue of the child's understanding *belief* with that of the child's developing a general theory of representation. attributing an explicit theory of decoupling to the child. What is attributed to "the child" is a representational system that makes available a PA-like notion of *belief*. This conceptual primitive can then be accessed by the child in, for example, learning more about beliefs, including learning about how some situations defeat the normally benign process of belief formation. We see no more reason to suppose that "meta-representation" *in Perner's sense* plays a major role in the preschool acquisition of ToM knowledge than to suppose that "metalinguistic" knowledge underwrites the child's acquisition of language.[5]

ToMM: *A theory of the capacity to acquire a theory of mind*

According to the Gopnik-Wellman view, the preschooler is at various times a "drive theorist", a "copy theorist", then a "representational theorist". Additionally, again according to these writers, the child has to patch her theories with a number of *ad hoc* "auxiliary hypotheses". According to Perner's view, the preschooler is a "behaviorist", then a "situation theorist", and then a "representational theorist" who, like the Gopnik-Wellman child, also has to employ a number of *ad hoc* "strategies" along the way.[6] We have been developing an alternative to the notion that the preschooler somehow discovers a succession of theories in his attempt to understand the behavior of agents. This alternative assumes much more continuity in the child's development. We view the changes in the child's behavior as stemming from incremental increases in the child's problem solving *abilities* rather than in radical changes in basic conceptual competence or structural knowledge (e.g., Leslie, 1987, in press; Leslie & Thaiss, 1992). Our framework addresses ToM

---

[5]By "metalinguistic" knowledge, psychologists mean, for example, knowledge that "the" is a word or that "bed" is a smaller word than "butterfly".

[6] *Ad hoc* supplements are required by these writers in order to explain various aspects of the child's performance. However, the "ad hocness" is attributed by these writers to the child! In the absence of independent evidence for each of the supplements, rather than regarding these as patches applied by the child to her own views, they are better regarded as *ad hoc* patches applied to the Gopnik-Wellman-Perner accounts.

development from the point of view of an information processing system. Some theory-theorists try to deal with information processing questions as if they formed a competing approach to theory-theories (e.g. Gopnik, 1993; Gopnik & Wellman, this volume). In fact information processing questions must be addressed, unless one supposes—*credo quia absurdum?*—that discovering and applying theories somehow does not require the processing of information (see our remarks in Leslie, German & Happé, 1993).

According to our alternative view, the critical factor in understanding our capacity to acquire theory of mind is a specialized, domain specific processing mechanism that employs a pre-structured representational system. We call this mechanism, **ToMM**, and the representational system it employs, the "metarepresentation" (or sometimes simply the "M-representation"). **ToMM** is a mechanism specialized by adaptive evolution for the task of interpreting agents' behavior in terms of propositional attitudes. As we saw earlier, the M-representation expresses four kinds of information: it identifies an Agent, an attitude, an anchoring aspect of reality, and a fictional state of affairs, such that the Agent holds the given attitude to the truth of the given fictional state of affairs in respect of the anchoring aspect of reality. For example, **mother pretends (of) this banana (that) "it is a telephone"**.

The child understands that mother is pretending that a particular real banana is a telephone (Leslie, 1987). This is one reason a representation with the above structure is required. Pretence must be distinguished from a number of other things with which theorists sometimes confuse it. For example, Piaget (1951) proposed a "symbolic play" interpretation wherein the banana *stands for* a telephone; however, mother does not assert that this banana stands for a telephone. She does something different: she pretends that this banana *is* a telephone. Nor is mother understood as emitting crazy "banana-telephone" behavior; mother pretends precisely that *this* banana is a telephone—neither that some unspecified banana nor that bananas in general are telephones, but that this very banana *here and now* is a telephone (see the discussion of opacity and decoupling in Leslie, in press*b*). The child is perfectly well able to understand that mother is simultaneously pretending that one empty cup is full of tea and another empty cup (which she had just pretended to empty) is empty. If the fictional content of pretend was not related via the Agent to the real world (by way of the relation '*pretends true of*') then a two-way relation, **PRETENDS (Agent, fiction)**, would suffice. As the facts stand, and as Leslie (1987) pointed out, 'pretends' is, and is understood as, a three-way relation, **PRETENDS (Agent, reality, fiction)**. Of course, exactly this M-representational structure is also required for understanding the relation '*believes*'.

A device like **ToMM** equips the child to process the behavior of agents in such a way that the effects of fictional states of affairs on actual behavior are made sense of—via the agent's attitude to the truth of the fiction. This does not require the child to conceptualize a mind stocked or a head filled with representations. It leaves the RTM to be acquired culturally, if at all, as a theory of what propositional attitudes "really are", in much the same way, we assume, that the atomic theory of matter is acquired as a theory of what substances really are.

The functioning of **ToMM** is evident at least from the time the child can understand pretence (between 18 and 24 months of age). As we saw, there is no available evidence to support a subsequent radical theory shift at four years. What *is* well established is that there is a shift in per-

formance on standard false belief tasks. Before we explain this change in behavior in terms of the removal of a conceptual deficit, we should be able to rule out limitations in whatever general problem solving resources are required by standard false belief tasks. Admittedly, this is not easy to do. We do not possess an adequate theory of such resources nor even a useful task analysis. What we can do is design tasks with which, by way of an identical task structure, we try to reproduce the general problem solving demands of false belief tasks while varying the specific conceptual content. We can then look to see if three-year-olds pass the control task. If they do, it would argue against general problem solving demands as the cause of their failure on false belief and thus strengthen the conceptual deficit case. The out-of-date photographs tasks fit the bill nicely having an identical task structure with a different conceptual content.[7] In the case of autistic children—most of whom pass the photographs "control" task while failing false belief—a general problem solving limitation is most unlikely to be the cause of their false belief failure (Leslie & Thaiss, 1992). By contrast, normal three-year-olds fail both the photographs "control" tasks and false belief (Zaitchik, 1990). So, in the case of the three-year-old, we cannot rule out that their failure on standard false belief tasks is due to general performance limitations. The evidence, then, that, according to Gopnik and Wellman, forms the foundation of the RTM account—the convergence of performance patterns across a number of tasks—is susceptible to a straightforward alternative interpretation, namely, that all those tasks make a set of performance demands that three-year-olds cannot meet.

Recently, we have begun to characterize a general processing component, the "Selection Processor" (**SP**), which we believe plays a role in a number of tasks, e.g., false photographs and false maps as well as standard false belief. We do not argue these claims here for lack of space but refer the interested reader to Leslie (in pressb), Leslie & Roth (1993), Leslie & Thaiss (1992), Roth & Leslie (in preparation), Surian & Leslie (in preparation). Here is the general idea. **SP** performs a species of "executive" function, inhibiting a pre-potent inferential response and selecting the relevant substitute premise. This mechanism, like many other "executive functions", shows a gradual increase in functionality during the preschool period. Some belief tasks (as well as some other theory of mind tasks like understanding pretence) plausibly do not require this general component or stress it less, and, in these cases, good performance is seen in three-year-olds (e.g., Mitchell & Lacohée, 1991; Wellman & Bartsch, 1988; Roth & Leslie, 1991; Zaitchik, 1991). The three-year-olds' difficulty, then, is due to the limitations of this general component which is stressed by a range of tasks, regardless of whether or not agents and attitudes are involved. Meanwhile, in the normal three-year-old, **ToMM** is intact.

The autistic child, by contrast, shows poor performance on a wider range of belief reasoning tasks, even compared with Down's syndrome children, with other handicapped groups, and with three-year-olds (e.g. Baron-Cohen, 1991c; Baron-Cohen, Leslie, & Frith, 1985; Leslie & Frith, 1988; Roth & Leslie, 1991). This disability is all the more striking alongside the excellent performance autistic children show on out-of-date photographs, maps and drawings tasks (Charman & Baron-

---

[7]"Appearance-reality" tasks are sometimes claimed to play the role of ruling out general problem solving limitations. But, as Perner, Leekam & Wimmer (1987) pointed out, such tasks are simply procedural variations on the standard "Smarties" false belief task. These tasks share both conceptual content and critical features of task structure.

Cohen, 1992; Leekam & Perner, 1991; Leslie & Thaiss, 1992). Autistic success on these latter tasks rules out a whole class of explanation for their failure on false belief in terms of general impairments in executive functioning, attending, lack of cooperation, working memory, simulation, poor motivation, and so on. The observed pattern of ability and disability can be succinctly explained on the assumption of a relatively intact **SP** together with an impaired **ToMM**—a mirror-image of the normal 3-year-old. Figure 2 summarizes the **ToMM-SP** model of normal and abnormal development.

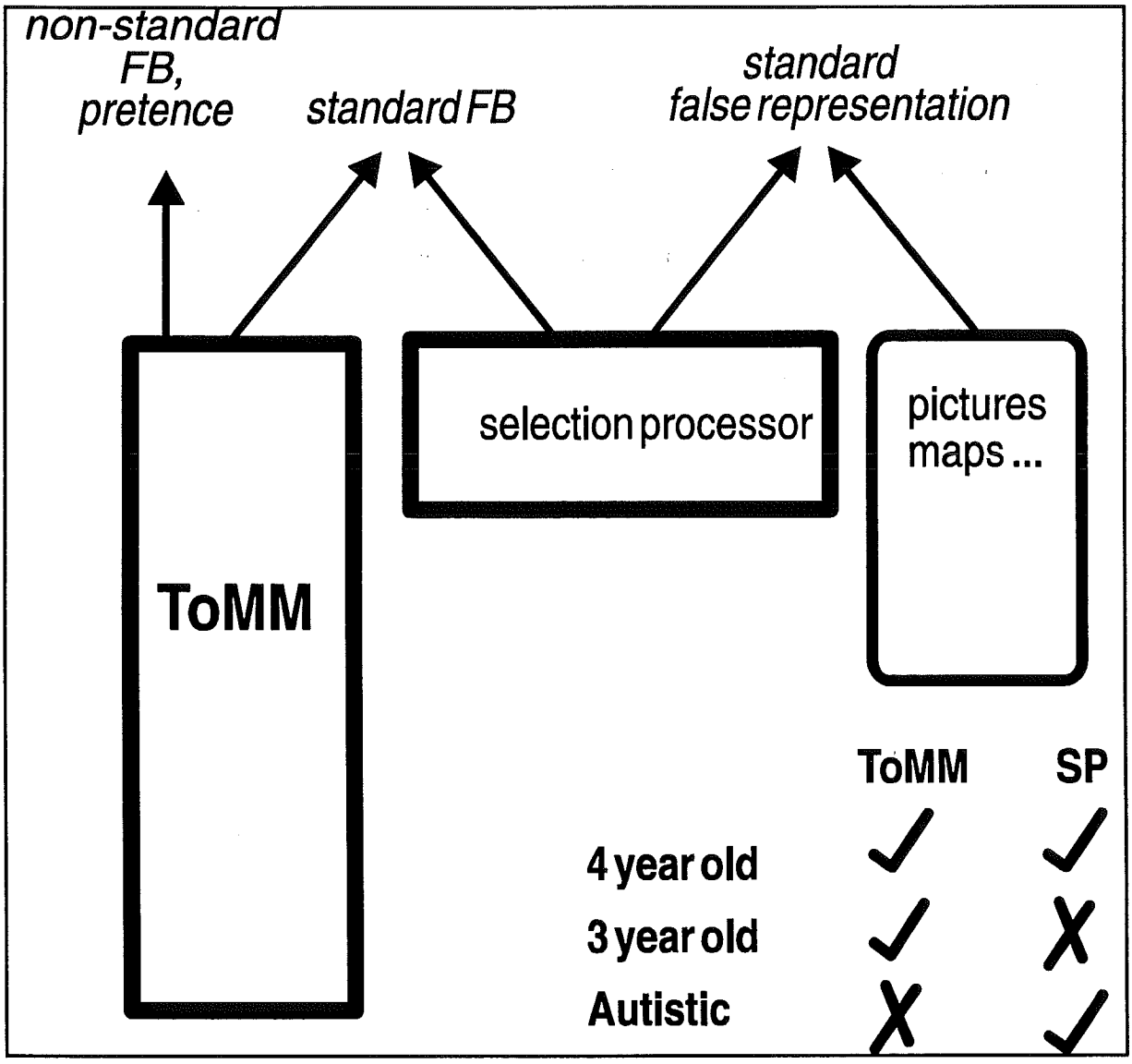

Figure 2. Schematic of the **ToMM-SP** model of normal and abnormal development, showing how normal 3-year-olds and autistic children are theoretical mirror-images of one another (after Leslie & Thaiss, 1992).

ToMM-SP *and Fodor's heuristic model*

There are a number of points of similarity between the **ToMM-SP** model and proposals made recently by Fodor (1992, reprinted in this volume). Fodor also argues for theory of mind competence in the three-year-old and also proposes a performance limitation to explain their failure on standard false belief tasks. Both theories assume that the preschooler has a *normative* concept of belief, that is, that the preschooler assumes that beliefs *ought* to be true (but sometimes they may be false). The theories differ on the role of this normativity. In the **ToMM-SP** model, the chief effect is to create a pre-potent response when it comes to inferring the content of someone's belief: the preschooler (like the adult) has a tendency to assume a belief content that he considers *true*. When faced with a false belief situation, the preschooler has to inhibit this pre-potent inferential response and instead select a specific, and now counterfactual, situation the other person was exposed to and enter this as the premise for the belief content inference. Perhaps preschoolers also assume that photographs are normatively true. In any case, they will have to select the appropriate, and now counterfactual, situation to which the camera was exposed and resist a competing representation of current reality, before making the correct photograph content inference. For these tasks, the "executive" functions of inhibiting and selecting call upon **SP** whose services are only tenuously available to three-year-olds.

In Fodor's theory, the three-year-old employs a *desire* based heuristic for predicting behavior. Only when this desire heuristic fails will the three-year-old try to infer the other person's belief. The four-year-old, with greater processing resources available, calculates the other person's belief routinely, regardless of how the desire based heuristic turns out. The three-year-old's desire based heuristic fails when there is an ambiguous prediction of behavior that will satisfy the agent's desire. Fodor gives examples of situations where an ambiguity in the object of desire should spur the three-year-old into calculating agent's belief. For example, when some chocolate is moved from one location and split between two new locations, a protagonist can satisfy desire for chocolate by searching in either of the two locations. This ambiguity should lead the three-year-old to calculate the protagonist's belief. The child will then realize that the protagonist still believes the chocolate is in its original location and predict behavior from belief. Unfortunately, three-year-olds do not perform better in this two location version of the standard false belief task (German, unpublished).

We think, however, that the particular examples Fodor gives are not, in fact, particularly good tests of his own hypothesis. In the above example, the child, not unreasonably, regards search in two locations as a single, unambiguous action. Surian and Leslie (in preparation) have devised a false belief task that more surely involves ambiguity in the object of desire. This task involves a protagonist asking for his "favorite pencil" and describing it as being "broken". The task is designed such that the only way the child has of determining which of four pencils is the protagonist's favorite is to track which pencil the protagonist *thinks* is broken. It so happens that there are *three* pencils which, unbeknownst to the protagonist, are broken now at the time of his request. The one pencil that was formerly broken, and which the protagonist knew about at the time, has, by the time of the request, unbeknownst to the protagonist, been sharpened. When the child looks for a pencil which is "broken", he is faced with three broken pencils. Any one of these could be the favorite if all the child has to go on is the protagonist's description. If, however, the child calculates the protagonist's (now false) belief in issuing that description, then a unique "favorite" will be identified. In fact, in

this situation, despite its complexity, the three-year-old performs significantly better than in a parallel scenario which lacks the element of ambiguity. This result provides support for Fodor's hypothesis.

Fodor's model assumes that, *if only she would try*, the three-year-old will have no difficulty in calculating correctly the content of a false belief. As far as we can see, this assumption is not correct. For example, three-year-olds still have difficulty in many false belief scenarios even when, instead of being asked to predict behavior, they are explicitly instructed to calculate belief. According to the **ToMM-SP** model, three-year-olds routinely calculate beliefs. The trouble is that routine calculation, for reasons outlined earlier, produces the wrong content for beliefs in standard false belief tasks. However, task structure can be altered in ways which help the three-year-old's weak **SP**. Ambiguity-of-desire and desire-anomalous-behavior are examples of factors which can function to promote the inhibition of the normative assumption. Leslie (in press*b*) discusses these issues further.

Whatever the details of the limiting performance factors turn out to be, one thing is becoming increasingly clear. The standard false belief tasks do not directly test the presence or absence of belief knowledge. What these tasks directly test is the ability to calculate, and calculate correctly, the contents of beliefs. The available evidence suggests that autistic children are impaired in their structural knowledge of belief, while three-year-olds have a limited ability in calculating belief contents. Neither the **ToMM-SP** model nor Fodor's heuristic model postulate a series of grand conceptual reorganizations; instead both rely upon the more modest assumption that the processing mechanisms of preschoolers gradually increase in efficiency.

The state of debate

The simulationist challenge has raised interesting questions about the nature of the action planning system. For example, what is the relation between generating goals and sub-goals in the course of planning an action and understanding the notion of goal-directed behavior? Do they both depend upon the same single underlying ability as simulationists would have it? Or are there two independent, non-interacting psychological entities? Or do independent entities, corresponding to goal-related knowledge and goal-related ability, exchange information? We know very little about such matters, but the openness of the questions underlines the seriousness of Fodor's suggestion (cited in SN1, pp. 47–48) that the action planning system might employ ToM-specialized knowledge structures.

SN2 argue that cognitive penetrability is the key issue dividing simulation and theory-theory views. We agree but have discussed this in terms of the contrast between the knowledge *together with* ability assumption of theory-theory versus the ability *only* assumption of simulation theory. The facts of development in the ToM domain require us to address the fundamental explanatory question concerning how it is young children are able to learn anything at all about mental states. We believe that the radical *ability only* version of simulation is implausible and, when formulated carefully, not

capable of addressing the fundamental question of learnability. We have criticized some existing versions of theory-theory for also failing to address this fundamental question, for ignoring the role of growing abilities, and for coming up with a new theory-theory for every change in the child's surface behavior.

When looked at in terms of its scope, the **ToMM** version of theory-theory is the best available account of both normal preschool development of theory of mind and of its abnormal development in the syndrome of childhood autism. The theory of **ToMM** answers some fundamental questions about the nature of the constraints upon early "hypothesis formation" in this domain. These constraints constitute a learning mechanism that allows normal development to begin very early and to proceed rapidly and uniformly. The theory makes correct predictions about both the detailed form of autistic impairment in this domain and the spared problem solving abilities on tasks which are close analogues of those on which autistic children show impairment. To the extent that simulation plays a role as a ToM ability, metarepresentation is required to organize and control the simulation process. In the case of simulation too, **ToMM** plays a fundamental ontogenetic role. Finally, together with an additional mechanism, **SP**, for which there is independent evidence, the model provides a way to begin to characterize the *knowledge* versus *ability* distinction for this domain. In this regard, it appears that preschool development of belief understanding is largely the result of increasing ability to use existing competence, while the abnormal development seen in autism results from impaired structural knowledge.[8]

---

[8] The knowledge-ability distinction can be drawn at different grains. So with increasingly sophisticated modelling of an information processing system, what appeared at one time to be simply knowledge, say, turns out to include an ability component as well. The converse may also happen. With more detailed models, what appeared to be simply an ability may turn out to have a knowledge component (e.g., the action planning system may employ knowledge). We do not want to rule out then that future work on the theory of **ToMM** will show that autistic impairment stems partly or wholly from an impaired **ToMM**-*internal* ability which then interferes with the deployment of **ToMM**-internal knowledge. See Leslie and Frith (1990) for further discussion of possible impairments to **ToMM**.

References

Baldwin, D.A. (in press) Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language.*

Baron-Cohen, S. (1991) The development of a theory of mind in autism: Deviance and delay? In M. Konstantareas and J. Beitchman (Eds), *Psychiatric Clinics of North America*, Special issue on *Pervasive Developmental Disorders*. pp. 33–51. (Pennsylvania: Saunders).

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

Baron-Cohen, S., Leslie, A.M. & Frith, U. (1986) Mechanical, behavioural and Intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology*, **4**, 113–125.

Baron-Cohen, S., Spitz, A., &Cross, P. (1993). Do children with autism recognize surprise? A research note. *Cognition and Emotion*, **7**, 507–516.

Blair, R.J.R. (unpublished) Role taking, empathy and the moral/conventional distinction. MRC Cognitive Development Unit, London.

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

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.

Chomsky, N.A. (1965) *Aspects of the Theory of Syntax*. (Cambridge, MA: MIT Press).

Chomsky, N.A. (1975) *Reflections on Language*. (New York: Pantheon).

Chomsky, N.A. (1988) Language and Problems of Knowledge: The Managua Lectures. (Cambridge, Mass.: MIT Press).

Curry, G. (unpublished). Simulation-theory, theory-theory and the evidence from autism. Department of Philosophy, University of Otago.

Fodor, J.A. (1992) A theory of the child's theory of mind. *Cognition*, **44**, 283–296.

Frith, U. (1989) *Autism: Explaining the Enigma*. (Oxford: Blackwell).

Frith, U., Morton, J., & Leslie, A.M. (1991) The cognitive basis of a biological disorder: Autism. *Trends in Neurosciences*, **14**, 433–438.

Goldman, A.I. (1993) Functionalism, the theory-theory and phenomenology. (Author's Response) *Behavioral and Brain Sciences*, 16, 101–108.

Gopnik, A. (1993) How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, **16**, 1–14.

Gopnik, A. (1993b) Theories and illusions. (Author's Response). *Behavioral and Brain Sciences*, 16, 90–100.

Grice, H.P. (1957) Meaning. *Philosophical Review*, **LXVI**, 377–388.

Harris, P.L. & Kavanaugh, R. (in press) The comprehension of pretense by young children. *Monographs of the Society for Research in Child Development.*

Karmiloff-Smith, A. (1988) The child is a theoretician, not an inductivist. *Mind and Language*, **3**, 183–195.

Leekam, S., & Perner, J. (1991) Does the autistic child have a "metarepresentational" deficit? *Cognition*, **40**, 203–218.

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

Leslie, A.M. (1988) The necessity of illusion: Perception and thought in infancy. In L. Weiskrantz (ed.), *Thought without Language*. pp. 185–210. (Oxford: Oxford Science Publications).

Leslie, A.M. (1992) Autism and the "Theory of Mind" module. *Current Directions in Psychological Science*, **1**, 18–21.

Leslie, A.M. (in press*a*) A theory of Agency. In (Ed.) A. Premack, *Causal cognition: A multidisciplinary debate*. (Oxford: Oxford University Press), in press.

Leslie, A.M. (in press*b*) *Pretending* and *believing*: Issues in the theory of **ToMM**. *Cognition*, in press.

Leslie, A.M. (in press*c*) **ToMM**, **ToBy**, and Agency: Core architecture and domain specificity. In L. Hirschfeld and S. Gelman (Eds.), *Mapping the mind: Domain specificity and cultural knowledge*. (New York: Cambridge University Press), in press.

Leslie, A.M., & Frith, U. (1988) Autistic children's understanding of seeing, knowing and believing. *British Journal of Developmental Psychology*, **6**, 315–324.

Leslie, A.M., & Frith, U. (1990) Prospects for a cognitive neuropsychology of autism: Hobson's choice. *Psychological Review*, **97**, 122–131.

Leslie, A.M., German, T.P., & Happé, F.G. (1993) Even a theory-theory needs information processing: **ToMM**, an alternative theory-theory of the child's theory of mind. Commentary. *Behavioral and Brain Sciences*, **16**, 56–57.

Leslie, A.M., & Roth, D. (1993) What autism teaches us about metarepresentation. In (Eds) S. Baron–Cohen, H. Tager-Flusberg, and D. Cohen, *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.

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

Mitchell, P. & Lacohée, H. (1991) Children's early understanding of false belief. *Cognition*, **39**, 107–127.

Ozonoff, S., Pennington, B.F., & Rogers, S.J. (1991) Executive function deficits in high-functioning autistic individuals: Relationship to theory of mind. *Journal of Child Psychology and Psychiatry*, **32**, 1081–1105.

Perner, J. (1991) *Understanding the Representational Mind.* (Cambridge, Mass.: MIT Press).

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

Piaget, J. (1962). *Play, dreams and imitation in childhood.* London: Routledge & Kegan Paul.

Pylyshyn, Z.W. (1978) When is attribution of beliefs justified? *The Behavioural and Brain Sciences*, **1**, 592–593.

Pylyshyn, Z.W. (1984) *Computation and Cognition: Toward a foundation for cognitive science.* Cambridge, MA: MIT Press–Bradford.

Roth, D., & Leslie, A.M. (1991) The recognition of attitude conveyed by utterance: A study of preschool and autistic children. *British Journal of Developmental Psychology*, **9**, 315–330. Reprinted in (Eds) G.E. Butterworth, P.L. Harris, A.M. Leslie and H.M. Wellman, *Perspectives on the Child's Theory of Mind.* pp 315–330. (Oxford: Oxford University Press) 1991.

Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991) The `windows task' as a measure of strategic deception in preschoolers and autistic subjects. *British Journal of Developmental Psychology*, **9**, 331–349. Reprinted in (eds) G. Butterworth, P.L. Harris, A.M. Leslie and H.M. Wellman, *Perspectives on the Child's Theory of Mind.* (Oxford: Oxford University Press) 1991.

Shah, A. (1988). *Visuo-spatial islets of abilities and intellectual functioning in autism.* Unpubl. Ph.D. Thesis, University of London.

Sodian, B. & Frith, U. (1992) Deception and sabotage in autistic, retarded and normal children. *Journal of Child Psychology and Psychiatry*, **33**, 591–605.

Stich, S., & Nichols, S, (1992) Folk psychology: Simulation or tacit theory. *Mind & Language*, **7**, 35–71.

Wellman, H.M., & Bartsch, K. (1988) Young children's reasoning about beliefs. *Cognition*, **30**, 239–277.

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

Zaitchik, D. (1991) Is only seeing really believing?: Sources of true belief in the false belief task. *Cognitive Development*, **6**, 91–103.