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.