Mind & Language, ISSN: 0268-1064
Vol. 13. No. 3 September 1998, pp 323–340.

Mindreading: Mental State Ascription and Cognitive Architecture

JOSEPH L. HERNANDEZ CRUZ

Abstract: The debate between the theory-theory and simulation has largely ignored issues of cognitive architecture. In the philosophy of psychology, cognition as symbol manipulation is the orthodoxy. The challenge from connectionism, however, has attracted vigorous and renewed interest. In this paper I adopt connectionism as the antecedent of a conditional: If connectionism is the correct account of cognitive architecture, then the simulation theory should be preferred over the theory-theory. I use both developmental evidence and constraints on explanation in psychology to support this claim.

1. Introduction

Our ability to explain one another's actions by appeal to inner mental states is distinctive for its sophistication, dynamic sensitivity and predictive accuracy. You are confident that I am crying because I am *sad*, that Clinton hesitates because he is *uncertain*, or that Smith is dialling long distance because he *believes* that Brown is in Barcelona. These explanations appeal to the mental states of sadness, uncertainty and belief respectively. We hypothesize mental states inside someone's head to explain her actions and, except for poetry and metaphor, we think it is a little odd to expound on the hopes, beliefs and desires of inanimate objects.[1] For intentional entities, a folk psychology that posits mental states has much to recommend it.

There has been an explosion of interest in this topic by cognitive scientists of all stripes under the rubric of theory of mind studies. The fundamental issue is, *how do we ascribe mental states, and how is it that we are so good at it?*

A shorter version of this paper was presented at the 1996 meeting of the Pacific Division APA and at the 1996 meeting of the Society for Philosophy and Psychology. I wish to thank audiences at these two meetings. In addition, I am especially indebted to Alvin Goldman, Terry Horgan, Jack Lyons, Joel Pust, Diana Raffman, Neil Tennant and to audiences at the University of Arizona and the Ohio State University for comments and discussion of this topic. Finally, I also thank two anonymous referees from this journal.

Address for correspondence: Hampshire College, Cognitive Science, Adele Simmons Hall, Amherst, MA 01002, USA.
Email: cruz@hampshire.edu.

[1] Some find it natural to extend the intentional stance to inanimate objects, so long as they are properly representational. See Dennett, 1987, 1995.

2. *The Theory-Theory and the Simulation Theory*

The debate over how we ascribe mental states to other cognizers has, for the most part, ranged over the limits of the conclusions one can draw from the experimental data resulting from studies on how children develop this capacity. Theory-theorists hold that adults possess a theory that somehow describes generalizations mediating behaviour or observable characteristics and mental states. Observed behaviour plus background knowledge is the input to this theory, while a mental state attribution is the output. The output is used by the ascriber to explain intentional action, to generate plans that involve the target and to predict future behaviour by the target. Of course, the theory is tacit; these transactions typically go on below the level of conscious awareness.

Prominent versions of the theory-theory claim that between the ages of 3 and 4 years, toddlers construct and refine the theory of mind (in the psychological literature, cf. Astington et al., 1988; Wellman, 1990; Wellman and Woolley, 1990; Astington and Gopnik, 1991; Gopnik and Wellman, 1992). The most dramatic innovation in the development of the theory occurs as children adopt a representational view of the mind where they begin to understand that their own mental states may differ from other's given differences in perceptual access and perspective (Gopnik, 1993). By the age of three and a half, the theory of mind allegedly begins to allow for these cases. The theory might be innate or learned, and might be modular (strongly or weakly) or part of the wider web of general folk theorizing about the world. The geography of the debate in psychology emerges nicely by categorizing proposals according to their answers to these questions.

I will not pursue in detail the theoretical and empirical considerations that motivate the array of research programmes on the theory-theory.[2] The idea is that, more or less in the same way that we have a folk theory of the behaviour of middle-sized objects, we have a theory of how to make sense of the activities of other sentient beings in terms of inner states. This is not to say that there aren't significant differences between, say, a folk psychology and a folk physics. The crucial common ground is that, in ways specified by the proponents of the view, both are properly theories.

The theory-theory is widely held in developmental psychology and cognitive anthropology. Some philosophers and cognitive scientists, however, have entered into the debate in part on the premise that the data from psychology does not conclusively support the theory-theory. A *simulation* account has been offered as an alternative and competitor to the theory-theory (Gordon, 1986; Heal, 1986; Goldman, 1989; Harris, 1989).

The simulationist's claim is that in order to ascribe mental states to others, people use the target's circumstances as make-believe inputs for themselves,

---

[2] For a first rate and up-to-date overview, see the introductions to the Davies and Stone collections, 1995a, 1995b.

and generate their own mental state off-line. This is the mental state that they then ascribe to the target. The input in the simulation theory may come from perceiving the target's circumstances or from a fund of background knowledge about the target or both. Jane Heal (1995) puts it this way: 'Suppose I am interested in predicting someone's action. . . . What I endeavor to do is to replicate or recreate his thinking. I place myself in what I take to be his initial state by imagining the world as it would appear from his point of view and I then deliberate, reason and reflect to see what decision emerges' (p. 47).

There are two commitments made in this passage that are typical of the simulation view. First, the simulationist must posit some manner of pretence that provides the input conditions for simulation. There must be some way that an agent can place herself in the shoes of a target of mental state ascription. Second, there must be a way to 'deliberate, reason and reflect' in such a way that the results are not action or belief guiding. According to the simulationist, the deliberative mechanisms in use here for mental state ascription are the same mechanisms used in the first person case. These double-duty mechanisms must be prevented from issuing in behaviourally efficacious results in the third person case. This is just what taking the mechanism 'off-line' amounts to. Theory-theorists must also endorse a pretence mechanism as the capacity for pretence is widely accepted in the developmental literature, completely apart from issues in the theory of mind (Fein, 1981; Bretherton, 1984). So, the single mechanism hypothesis is what is distinctive about the simulation account. One cognitive apparatus performs double duty.

The simulationist can concede that not all instances of mental state ascription are the result of simulation (Goldman, 1989). We may over time develop inductively based generalizations that allow us to assess without simulation new cases that are similar to old ones that we did simulate. The important point is that mental state ascription is fundamentally and first a simulative process. The simulation proposal gets some of its intuitive impetus from the insight that the most efficient and accurate guide to *other people's* mental state to behaviour transitions will be the cognitive mechanism that performs this function in one's own case. The simulationists have further argued that their account can adequately explain the psychological data; theory-theorists have attempted to rebut simulation by appeal to further studies (Nichols et al., 1996), and even now there is likely a study in progress that will, according to its authors, definitively refute the theory-theory or the simulation view.

Here I propose to take a different tack. I shall argue that commitments to views on the nature of representation and cognitive architecture can have an important bearing on our consideration of issues in the theory of mind. My strategy illustrates another avenue of participation in the theory of mind debate. The philosophy of psychology has recently seen a revival of careful consideration of how specific views on representation influence psychological science (Clark, 1993; Von Eckhardt, 1993; Macdonald and Macdonald,

1995b). The appearance of connectionist accounts of cognition has pressed these issues to the fore (Macdonald and Macdonald, 1995a).

I will take the recent emergence and growing acceptance of connectionist accounts of cognition as a starting point. That connectionism has a significant and apparently burgeoning influence on psychological theorizing shall be enough to generate a conditional: If connectionism is the proper global account of cognition, do our theoretical commitments here speak to the theory of mind debate? My answer is affirmative and in favour of simulation.

3. *Connectionism*

I turn to a brief rehearsal of the connectionist's claims about the architecture of cognition. This will lay the groundwork for the suggestion that representational commitments do speak to the theory of mind debate.

Though many of the essential components and insights of connectionism have been around for some time (McCulloch and Pitts, 1943; Hebb, 1949; Rosenblatt, 1962; Minsky and Papert, 1969), it is only in the last decade that parallel distributed processing accounts of cognitive activity have posed a serious and pressing challenge to the view of the mind as a serial symbol manipulator (Rumelhart and McClelland, 1986; Smolensky, 1988; Clark, 1989). Connectionist architectures are characterized by a set of simple nodes that instantiate transfer functions. These transfer functions accept input from the environment or other nodes, and produce output that is scaled depending on the inhibitory or excitatory nature of the links between the nodes. A network thus produces a total activation vector resulting from the weights of all the connections in the network. The total activation vector is then functionally associated with a particular input. Via various feedback rules—some mathematically complex, others not—the network can be trained to output a particular activation vector when given a particular input. Weights can be updated to produce the proper output in response to novel input, while still preserving the input–output mappings for old training data. The network thereby effects an important kind of learning.

Even a casual perusal of the current literature in the philosophy of mind and cognitive psychology will show that connectionism has made significant inroads into the way we think about cognition, broadly construed. There have been relatively fewer attempts, however, to link architectural questions with particular issues in psychology. Although the relationship between architectural considerations and cognitive psychology has been discussed, only a minority of working researchers have tried to exploit these insights. This is where the nascent field of cognitive neuroscience shines. Kosslyn and colleagues (Kosslyn et al., 1992), for example, have used considerations of connectionist architectures to argue that the visual system exploits two kinds of spatial relations—coordinate and categorical—to process objects. Of course, merely having a connectionist argument for a psychological theory is insufficient. Convergent data from other methodological domains will be

required. My argument below, then, should not be overstated. I claim that connectionism gives *one* source of evidence in favour of simulation. There is still much empirical and theoretical work to be done.

Logically, there are several possibilities. One might think that connectionism does not weigh in favour of the theory-theory or the simulation account in any way. After all, Paul Churchland (1989) advocates the theory-theory, only to complain later that it is not a very good theory and should be eliminated. The implication is that being a connectionist does not affect one's scruples regarding the theory of mind. The irrelevance thesis is explicitly adopted by Stich and Nichols (1992):

> The difficulties encountered by those who have sought to describe the rules or principles underlying our grammatical . . . abilities have convinced a growing number of theorists that our knowledge of these domains is not stored in the form of rules or principles. That conviction has been an important motive for the development of connectionist and other sorts of non-sentential and non-rule based models. But none of this should encourage an advocate of the off-line simulation theory. The dispute between connectionist models and rule-based models . . . is a dispute *among theory-theorists*. (p. 49, emphasis in original)

It is unclear how we should read the emphasized claim. It may very well be an accurate description of the literature that connectionist versus sentential accounts have been proposed as alternatives only among theory-theorists. This would not be a terribly interesting point. Stich and Nichols must be urging something more serious: That nothing about connectionism could make a difference in the theory-theory/simulation debate because the tension between connectionist and symbolic architectures occurs at a different level of inquiry than that of the theory-theory and simulation theory. According to Stich and Nichols, one must *first* decide whether a theory-theory or an off-line simulation account is to be endorsed for mental state ascription. If a theory-theory is endorsed, only then will the issue of deciding what manner of representation to embrace arise.

But is this a legitimate way of construing the dialectic? Suppose it turned out that one account or another of mental state ascription was suggested by features *specific and peculiar to the underlying architecture*? Would it suffice to react to this by claiming that the revelation is irrelevant because the choice of a theory-theory versus a simulation theory occurs before architectural issues need to be considered? It is surely against the spirit of the methodology of contemporary psychology to abstract away completely from architectural constraints at the level of primary psychological theorizing. Worries about implementation are part of the inspiration for connectionist accounts, so it would be odd if this commitment was abandoned when trying to arbitrate between the theory-theory and the simulation theory. And it is difficult

to see what Kosslyn's laboratory is up to if we disallow architectural considerations from serving as a source of constraint on psychological theories.

I will argue that specific features of cognitive architecture weigh in favour of one theory of mental state ascription, the simulation theory.

4. *One Network Realizations*

Let us look at an adaptation of connectionism that instantiates the *theory-theory* of mental state ascription. I will show that the most plausible attempt to instantiate the theory-theory in a connectionist network will look alarmingly like the simulation proposal. I conclude that there is a deep affinity between the claims of the simulationist and the architectural constraints of PDP.

One might think that connectionism is poorly situated to explain or instantiate a theory in any domain, let alone a theory of folk psychology. The worry is that, in the absence of an explicitly represented set of symbols, it is difficult to see where a theory—with its transition laws and sanctioned inferences—might reside. Presumably the theory is somehow represented in the weights and activations of the network. But are weights and activations properly construed as a theory? If not, the simulation argument would be much easier. Namely, suppose we are connectionists, and connectionism cannot handle theories. Therefore, connectionism and the theory-theory are incompatible and simulation emerges as the victor. Of course, this argument comes at too great a price. For folk physics or folk biology, the theory-theory is arguably the only game in town. If either folk physics (or quantum mechanics, for that matter) as construed as a theory or connectionism has to go, prospects look dim for connectionism. So, this quick argument must be avoided. My view is that connectionism can instantiate theories, but that—in the case of mental state ascription—there is reason to think it does not. The theory-theory is not ruled out straightaway if we endorse connectionism. We may yet have a theory of biology or geology or folk geology—if there is such a thing—and remain connectionist. How would the theory work in the specific case of folk psychology?

We may suppose that there will be a set of vectors in the weight space comprised by the network's activation and connections that represent a theory of mind. When the behaviour of another agent is introduced into the network as input, the activations are recalculated, generating an output. That output will be the ascription of a mental state. Theory-theorists must claim that there is one set of activations that mediate a child's first person mental states and behaviour, and a second, separate, set of activations that mediate the ascription of mental states to other cognizers' behaviour by instantiating a theory. The idea is that, given these two separate cognitive tasks, two sets of activations will be required.

Two separate computational tasks, then, must be performed by either a single network or two different networks: One task is first person mental

state to behaviour mediation, the other is third person mental state ascription. Consider first a one-network hypothesis. In the next section we will consider the two-network hypothesis. What does success in two different tasks amount to in a single connectionist network? In a fully distributed neural network each node of the network is *representationally significant* (Van Gelder, 1991). Any representational state the system is in is composed of *all* the nodes. It therefore makes no sense to talk about a theory instantiated in some discrete portion of a network. Specifically, it will *not* be the case in a single fully distributed network that some nodes and connections instantiate a theory of ascribing third person mental states while another portion of the network simultaneously mediates first person mental states to behavior. The very same nodes do the work of both; the differences must be accounted for by appeal to the activations of the very same pool of nodes at different times.

Not all connectionist networks are fully distributed. Various grades of *localist* representation are possible, with individual representations supervening on single nodes or small clusters of nodes. In Sejnowski and Rosenberg's familiar text-to-speech system NETtalk, each letter is locally represented by one of twenty-six dedicated units at the input level (Sejnowski and Rosenberg, 1987). In a theory-theory network with local representations, some portion of the network might represent a theory of mind for third person ascription, while another non-overlapping portion of the network might represent the control structure for first person mental state to behaviour generation. Contrary to this possibility, the present argument will turn in part on treating the theory-theory network as fully distributed (or at least as not radically localist).

What are the merits of this assumption? In the connectionist literature, the move from local to distributed representation has two motivations (Clark, 1989; Feldman, 1989). First, distributed representation will afford fault tolerance or graceful degradation. In one sense of graceful degradation, this is the property that allows a system to behave coherently when offered incomplete or somewhat inaccurate input. Where representation is local, missing or inaccurate data would result in catastrophic failure because what would be missing or erroneous is an entire representation rather than a part of a representation. Where a representation is distributed, one node mis-firing or not firing at all will make relatively less difference to the overall performance of the system.

Second, distributed representation will allow for flexible generalization based on fine-grained similarities of representation. In a localist architecture, either the activation of the representation is present or not. In a distributed architecture, representations appear in degrees depending on similarity of the activity pattern to past activity patterns. Both the theory-theory and first person mental state to behaviour mediation appear to robustly exhibit the characteristics that motivate connectionists toward distributed representation. The theory-theory offers coherent output even where some of the input data is erroneous or missing. We can judge that someone is in the mental state of distress if we only see his tears while failing to notice his

trembling shoulders. Further, part of the impressiveness of the theory-theory is that it is subtly sensitive to the gradations of circumstances when producing mental state ascriptions, and can accurately attribute a mental state in novel interpersonal contexts (Fodor, 1987).

Finally, let us be clear on what is being assumed by stipulating a fully distributed network. I am not suggesting that entire brains or the hemispheres or even the lobes are single fully distributed networks. My claim is that within a single module, full distribution of representation is the most promising architecture. So, in whatever functional modules implicated in the theory of mind literature, I will presume that these are fully distributed connectionist networks. As a consequence, in the one-network hypothesis there will be a network that achieves both first person behaviour and third person attribution, and the representations that mediate the two tasks will each be distributed over the entire network.

Our one network, then, manages both the first person and third person mental state tasks. The network will require some control mechanism to determine which task to perform at any given moment. That is, there needs to be some representation of the difference between first person and third person processing. A first person-representing sub-component may appear as a single additional activity pattern at the level of input, or may appear in some more nuanced way across the network activation.[3] For simplicity, consider the case where the first person status of the processing is encoded as a single additional activity pattern. Borrowing some terminology from the philosophy of language, I will call these *de se* input vectors, as they are essentially first person indexical (Lewis, 1979; Perry, 1979). The *de se* vector will be the portion of the network's input activation that indicates to the single network that it is the first person that is the source of the mental state conditions.

The *de se* vector will not need to generate a dramatic effect on network activity. Its role is to tag an instance of processing as first person processing so that 'down stream', during later cognition, proper motivation and behaviour will result. In fact, it is important that the *de se* vector *not* have a massive influence on processing. This is because mental state to behaviour inferences are closely mirrored in first person and third person processing. For example, from the mental state of hunger we may, *ceteris paribus*, infer food-finding plans in both the first person and third person case. This thesis is recognized by both camps of the theory-theory/simulation debate. Were it not for inferential similarity, the theory-theorist's case would be clear. She could point out that third person mental state ascriptions are nothing like the inferential chains that go on in the first person case, so simulation must be false. This argument, however, is rejected because other peoples' mental states and patterns of behaviour are judged to be much like our own (Grandy, 1973). The reason simulation even begins to look plausible is infer-

[3] The latter possibility was suggested by one of the anonymous referees for this journal.

ential similarity between first and third person mental state tasks. It is a similarity that must be respected in this theory-theory one-network scenario. If the *de se* vector had a large effect on processing, inferential similarity would be difficult to maintain.

Connectionist networks work on a principle of conservatism. Where two objects to be represented are similar, e.g. the shape of the letters *O* and *U*, the representational states of the network will be similar. More precisely, the difference in weights between the state representing the prototype of 'O' and the state representing the prototype of 'U' will be less than the difference in weights between the state representing the prototype of 'O' and the state representing the prototype of 'K'. A liability of this feature is that sometimes a network will mistake similar objects in, for instance, pattern recognition tasks. On the other hand, this conservatism allows the network to process noisy input so that an imperfect token of 'U' will still be processed as a 'U' as long as the figure is not more similar to some other letter form (Rumelhart and McClelland, 1986; Mozer, 1991).

In a single network, full distribution together with representational conservatism presents a problem for the theory-theory. Here is a pared-down version of the argument. Where the theory-theorist's hypothesis is instantiated in a single connectionist network (by the one-network hypothesis), the same nodes and weighted connections are responsible for a theory of third person mental states and for mediating first person mental states to behaviour. The difference between the two cognitive processes is due to what we have been calling the *de se* vectors. These are part of cognitive processing during those times of first person mediation between mental states and behavior. At other times, the activations that constitute the *de se* vectors will be absent. Where the *de se* vector's effect is slight, a connectionist realization of the theory-theory of mental state ascription will be in a computational condition very close to the condition where the system is generating behaviour from its own mental states. The third person mental state theory must therefore *be* the first person mental state network activation and weight with the *de se* vectors subtracted out. Given the principle of conservatism, the total activation resulting from the subtraction will be *very* similar to the total activation representing the first person state.

Unfortunately for the theory-theorist's case, the proposal is essentially the simulation account: Ascribing third person mental states is a matter of taking one's own mental state network off-line and running the usual process. The absence of the *de se* vector amounts to bringing the network 'off line'; that is, without the *de se* activation, the network will run without the information that this particular computational transformation is the first person case. Though it is possible to insist that this new vector is a theory while holding that the *de se* inclusive total vector is not, this is unprincipled. Recall that what is crucial about the simulation account is that the same functional machinery be used for both first person and third person mental state tasks. Where only one network is involved, the connectionist theory-theorist is drawn toward simulation.

I conclude that by the one-network hypothesis, a connectionist will be a simulationist. In the next section, the two-network hypothesis is investigated to determine whether or not the theory-theory can be rescued from within a connectionist framework.

5. *Two Network Realizations*

In the two-network hypothesis, one network will be responsible for first person mental state mediation, while a second separate network will be responsible for third person attribution. Given the considerations above, both will be fully distributed. Therefore, it is possible for one network to be a theory of mental state ascription, while the other is (potentially) a non-theoretical first person mental state mediating mechanism. This looks like a promising line for the connectionist theory-theorist.

I will investigate two sources of evidence that show that the two-network hypothesis cannot be right. Ultimately it will not do to hypothesize two distinct networks with two different sets of weights. First, we look to the psychological data that militates against the two-network hypothesis. Second, I will argue from principled requirements on explanation in psychology to show that the two-network hypothesis should be rejected.

The errors that toddlers make in mental state ascription are errors that are characterized by ascribing *their own* mental state to others rather than some other, incorrect, mental state. This point is illustrated by a well-known experiment by Astington and Gopnik (1991)—based on a paradigm due to Perner et al. (1987)—where 3-year-old children are presented with a sweet box. Upon inspection, the sweet box turns out to be full of pencils. When the children are asked what someone else will guess as the contents of the box, they report pencils, even though *they* thought that the box contained candy before it was opened. This seems to show that, if children are using a theory at all, they are using a theory that has access to first person mental states.

The error is even more dramatic in the original Perner paradigm. In his experiments, 3-year-olds are shown a character, Maxi, putting a piece of chocolate away in cupboard A. Maxi goes out to play, and his mother moves the chocolate to cupboard C. The young subjects of this experiment are privy to this plan. Maxi returns, craving chocolate, and the children are asked which cupboard Maxi will look in. Children respond statistically significantly with cupboard C, failing to attribute the false belief to Maxi. A control condition indicates that these subjects remember where Maxi put the chocolate. Moreover, *no* subject responded with cupboard B. If the first person and third person systems were completely modular, the error would be a mystery as there would be no explanation of why the child was seduced into predicting that Maxi would look in the cupboard that she, the child, believed the chocolate to be in. These results have been replicated many times over a variety of experimental paradigms.

We are faced with explaining the particular error that is made on the above false belief tasks. The children in the experiments respond to questions with their own mental state, showing that the process that is responsible for mental state ascription has access to first person mental states. But with two fully distributed networks, the only way two networks could have access to one another is by being *connected*. That would create one network, thereby landing us back in the one-network hypothesis. At least *prima facie*, then, the two-network hypothesis seems to be unacceptable on empirical grounds.

It may seem that this problem looms for orthodox, symbolic approaches to cognition and that the same empirical data would be equally effective against the theory-theory in more traditional architectural realizations. If this were so, connectionism would not be doing any of the work in the argument against the two-mechanism view; the data from psychology would argue for simulation. Happy as this result would be for the simulationist, the argument just does not go through with an orthodox architecture. This is because the orthodox view has an extra degree of explanatory freedom with which to account for the error in the false belief task. The additional freedom owes to what we shall call a data structure/process distinction that is absent from connectionism (Cummins, 1989, p. 154).

When an orthodox symbol theory-theorist attempts to explain the pattern of empirical results in mental state ascription, he has two items in his toolkit. He may appeal to the data structures encoded in the systems' memory or perceptual output and to the processes that take the data as arguments. Two processes may look to the same data pool for their arguments. For example, take *imagining* as one process, and *looking-at* as another. Naturally, these are too course-grained to do any real work in cognitive psychology and are meant for illustrative purposes only. So consider imagining something specific, like Socks the Cat. Alternatively, consider looking at Socks the Cat. The two processes are distinct, but there is no difficulty in holding that they compute over the same argument, namely, the Socks representation. There is nothing illicit in maintaining that all the difference between imagining Socks and looking at Socks is explained by appeal to the processes while holding the data constant. Another way of putting this, familiar to computer programmers, is that in traditional programming languages such as LISP there are variables and there are functions. The variable theta in the LISP function (SETQ counter θ) is the same theta as the theta in the function (PUSH θ accumulator), even though SETQ and PUSH are different processes and result in different computational states.

Returning to the theory of mind, suppose one wanted to explain a child's reporting her own belief when asked about the beliefs of another child who should have a false belief. One might claim that there is a process that mediates first person mental state computation and a second process that deploys the folk theory of mental state ascription. For toddlers, but not for adults, the data appropriate to the first process infects the second process. Infection here means erroneously contributing the first person data to the

third person process. As this proposal might go, there are two processes, but the error that children make is parsimoniously explained.

This is not a 'just so' story. Leslie offers this very account (Leslie and Thaiss, 1992). He argues that a standard false belief task such as the sweet box experiment requires two components. The more general component will be the Theory of Mind Module (ToMM). This is the theory of mind mechanism proper. The second component is a selection processor (SP), responsible for deciding which pool of information to draw from to determine input to ToMM:

> According to the model, 3-year-olds fail . . . standard false belief . . . tasks because SP is as yet poorly developed. They do not fail because they cannot remember the previous and now counterfactual state of affairs—control questions show otherwise—but because they fail to select the appropriate counterfactual or because reality intrudes. To succeed, the child must identify and select the right premise to enter into the inference process, resisting intrusion from other premises. (p. 247)

In Leslie's proposal, the input premises to the ToMM are separate from the ToMM. In orthodox accounts, this is a perfectly reasonable division of labour. So, with respect to my arguments, all bets are off if we start with the orthodox computational view.

It may seem that the orthodox approach contains valuable lessons for the connectionist theory-theorist, but it will not help to simply copy the architecture offered by Leslie. For Leslie, ToMM is one process, corresponding to the one-network approach. However, might there be available a thoroughly connectionist explanation that roughly mimics the orthodox approach in its use of two networks? In this proposal the psychological data that shows failure on the false belief task is accounted for by hypothesizing that the input vector reflecting first person states is erroneously input to the third person ascription network. This seems possible as long as two networks can potentially take the same vector as input.

While the suggestion in question could emulate the orthodox approach to taking into account the false-belief error, it will not suffice as a psychological explanation of the error. This is because it is viciously *ad hoc*. As we found above during the discussion of the role of the *de se* subvector in the one-network hypothesis, the two networks posited would have to be inferentially very similar because the generalizations that describe mental state transactions are largely the same in the first person and third person case. What would explain the identity of the inferences, given two networks? A connectionist could *stipulate* that the two inference mechanisms in the two networks are the same, but as Leslie himself, and Fodor and Pylyshyn, have argued, this is an unprincipled psychological explanation. The stipulation is the flaw.

Leslie views it as important that his theory-theory account be realized in an orthodox, symbolic architecture (Leslie, 1988). ToMM is a single process

that potentially receives two kinds of data. He insists that in order to mimic this explanation connectionistically:

> . . . One could attempt to construct [in this case, two] networks whose 'contents' had shared properties. On the other hand, it would be just as easy to construct networks whose 'contents' were arbitrarily different. There is nothing in connectionist architectures to prevent functionally distinct networks from differing arbitrarily. But this fact will deprive us of a principled explanation should it be the case that different 'mental spaces' are *always* related in their content. (p. 206)

The contents referred to in the passage are the inferences. Leslie therefore uses some of the very same psychological data reported here coupled with connectionism's inability to make the data/process distinction as an argument for the orthodox view. I have attempted elsewhere to fend off this anti-connectionist conclusion (Cruz, 1995).

What is at the heart of the issue is that the parsimony of a manoeuvre in cognitive psychology will depend on whether we start with a connectionist architecture or an orthodox architecture. Leslie's point and the argument strategy I appeal to above to refute the two-networks hypothesis is closely related to Fodor and Pylyshyn's admonition that connectionism cannot account for *inferential coherence*[4] (Fodor and Pylyshyn, 1988). As Fodor and Pylyshyn urge, '. . . inferences that are of similar logical type ought, pretty generally, to elicit correspondingly similar cognitive capacities. . . . This is because, according to the Classical account, this logically homogeneous class of inferences is carried out by a correspondingly homogeneous class of psychological mechanisms' (p. 47). The lesson we are to draw is that, for a genuine psychological explanation of the similarity between homogeneous inferences, a single process must be implicated. Orthodox computationalism meets this desideratum admirably by positing a single process working on a multiplicity of data. The problem with connectionism, according to Fodor and Pylyshyn, is that *'connectionists can equally model a mental life in which you get one of these inferences and not the other'*, even though the inferences are homogeneous (p. 130, emphasis in the original). This argument has the same structure as Leslie's, but Leslie casts it between networks rather than within a network. Leslie's argument is therefore shielded from prominent (and, to my mind, convincing) replies to Fodor and Pylyshyn (Smolensky, 1995).

My use of this argument holds that, even though Fodor and Pylyshyn's points with regard to a single network have met with an effective counter, they yet obtain with respect to multiple networks in just the way Leslie

---

[4] This characteristic was called *systematicity of inference* in the original article. The label *inferential coherence* later replaced it.

claims. For my purposes and on the same grounds, the argument shows that the two-network attempt to explain the false-belief task failure should be rejected. The developmental evidence indicates that the cognition responsible for the first person mediation between mental states and behaviour are part of the same cognitive apparatus that must embody the putative theory-theory. Again, if this were not so, the theory-theorist would owe us an account of why children in the false belief task make the particular misattribution that they do, namely, offering their own mental state. We are back, then, to a single network hypothesis: If a single fully distributed network runs two cognitive events, all the elements of that processing take place over all the nodes of that network.

A happy result of this theory-theorist cum one-network connectionist proposal is that it explains the data from cognitive development. When queried on the mental state of others, children at age three are unable to distinguish the mental states of others from their own. The cost of this success is that we are back to the argument scouted in Section 4; starting with connectionism, simulation looks very natural.

The argument of this section relies crucially on two pieces of information. The first (empirical) point is that children are prone to substitute first person mental state ascriptions in third person ascription tasks. This establishes that the two processes have intimate access to one another. The second (methodological) point is that the inferential coherence between first person mental state mediation and third person ascription makes positing two distinct networks *ad hoc*. Thus, a single network hypothesis, with its simulation-favourable consequences, is endorsed.

What of our other folk capacities, such as folk physics? It would be a radical conclusion to maintain that a single network was responsible for our judgements in all folk domains. So, my arguments must allow for a way to keep the folk physics network, for example, separate from the folk psychology network.[5] While the capacity to ascribe mental states seems to have access to the output of the folk physics network (for ascribing beliefs about the behaviour of middle-sized objects), there is no reason to view them as subsumed by one network. First, there is not a pattern of empirical results showing that children or adults systematically substitute folk physics judgements for folk psychology judgements. Thus, folk physics information appears more like data for the folk psychology network than output of the folk psychology network, in contrast to the first person output bias that children show. Second, folk physics and folk psychology are not inferentially coherent. The pattern of inferences in folk psychology is only distantly related to the pattern of inferences in folk physics. Thus, the methodological argument that makes a two-network—one first person, one third person—hypothesis suspect in folk psychology, does not argue against separating folk psychology and folk physics. Presumably, a similar point could be made

[5] This point is due to an anonymous referee for this journal.

about employing my argument to collapse any folk domain with folk psychology.

6. Conclusion

I have appealed to unique features of connectionist architectures and to developmental data in order to show that the simulation approach is favoured by connectionism. As the developmental data are crucial to the argument, it is yet open to the theory-theorist to dig in her heels and point out that the reported experiments give data on only one kind of ascription achievement. It remains possible to hold that other kinds of mental state ascription, for different mental states or in different contexts, would not show the pattern of empirical results necessary to yield my argument. Roughly, the empirical results that the argument relies on show that third person mental state ascriptions have access to first person mental state ascriptions. This is what is essential about children's failure on the false-belief task: they mistakenly ascribe their own beliefs.

Part of the problem for my case is that adults do not fail the false belief task. Other evidence must be marshalled to show that in adults mental state ascriptions reference first person mental states. There is some suggestive data available beyond children's failure on false beliefs. In the social psychology literature, Geis and Levy (1970) focus on subjects who achieve low scores on tests to determine their level of willingness to engage in Machiavellian manipulation of others. Their data shows that these low-scorers do worse than medium and high-scorers on predictions of what other test-takers will score on such tests. This is because they grossly underestimate the scores of other test-takers, instead predicting a score much closer to their own. This is plausibly viewed as a case of the low-scorers mis-attributing *their own* mental states to others (for discussion, see Mealey, 1995). By this interpretation, the low-scorers are victim to something akin to an adult version of failure in the false belief task.

Similarly, Dodge and Newman (1981) report evidence that shows that, in teenage boys, aggressiveness covaries with an over-attribution of aggression in others (see also, Nasby et al., 1979; Steinberg and Dodge, 1983; Waas, 1988). In one experimental setup, adolescents incarcerated for unacceptable aggressive behaviour reliably over-attribute aggressive intentions in neutral test scenarios compared to non-aggressive subjects. More recent studies have shown that the amount of over-attribution is correlated with the amount of aggressiveness manifest in the attributer, here gauged as a function of the seriousness of his crime (Dodge et al., 1990).

These experiments, though suggestive for the simulationist's case, are clearly not decisive. They do, however, move some way toward demonstrating the weaker claim that mental state attributers have direct access to and are influenced by their own mental states. This is the type of evidence that is required in an expansive range of cases to solidify my argument above.

The future research programme of simulationists must be aimed at further demonstrating the dramatic interference of first person mental states in third person attribution. The current data available, conjoined with architectural considerations of connectionism, makes the simulationist's case widely compelling and puts the burden of proof squarely on the shoulders of the theory-theorist.

The simulation account that results from connectionist considerations is only a *version* of simulation. The argument adduced above suggests that *not all* of the vectors used in the first person generation of mental states are used to simulate. During simulation, the *de se* vector will be absent from the input activation to the network. So, strictly speaking, the network employed is different, activation-wise, during simulation. In spite of this activation difference, the same machinery is in use for third person mental state ascriptions. It would be strange to hold that the same network was at some point a mental state to behavior mediating module, and then becomes a theory of mental state ascription simply by subtracting out the *de se* vectors at the level of input. The more natural conclusion is that the same mechanism, or a decisively similar one, is responsible for both first person mental state to behaviour generation, and third person mental state attribution. This is just what simulationists hold.

Cognitive Science
Hampshire College

References

Astington, J.W. and Gopnik, A. 1991: Theoretical Explanations of Children's Understanding of the Mind. *British Journal of Developmental Psychology*, 9, 7–31.

Astington, J.W., Harris, P.L. and Olson, D.R. (eds) 1988: *Developing Theories of Mind*. Cambridge University Press.

Bretherton, I. 1984: Representing the Social World in Symbolic Play: Reality and Fantasy. In I. Bretherton (ed.) *Symbolic Play*. New York: Academic Press, pp. 1–41.

Churchland, P. 1989: *A Neurcomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press.

Clark, A. 1989: *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge, MA: MIT Press.

Clark, A. 1993: *Associative Engines*. Cambridge, MA: MIT Press.

Cruz, J. 1995: Connectionists Don't Have To Pretend They Can't. Paper presented in Stony Brook, N.Y.: The Annual Meeting of the Society for Philosophy and Psychology.

Cummins, R. 1989: *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Davies, M. and Stone, T. (eds) 1995a: *Folk Psychology*. Oxford: Blackwell.

Davies, M. and Stone, T. (eds) 1995b: *Mental Simulation*. Oxford: Blackwell.

Dennett, D. 1987: *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. 1995: Do Animals Have Beliefs? In H. L. Roitblat and J.-A Meyer (eds), *Comparative Approaches to Cognitive Science*. Cambridge, MA: MIT Press, pp. 111–18.

Dodge, K.A. and Newman, J.P. 1981: Biased Decision-making Processes in Aggressive Boys. *Journal of Abnormal Psychology*, 90, 375–79.

Dodge, K.A., Price, J.M., Bachorowski, J.-A. and Newman, J. P. 1990: Hostile Attributional Biases in Severely Aggressive Adolescents. *Journal of Abnormal Psychology*, 99, 385–92.

Fein, G.G. 1981: Pretend Play: An Integrative Review. *Cognitive Development*, 52, 1095–118.

Feldman, J.A. 1989: Neural Representation of Conceptual Knowledge. In L. Nadel, L.A. Cooper, P. Culicover and R.M. Harnish (eds) *Neural Connections, Mental Computation*. Cambridge, MA: MIT Press, pp. 68–103.

Fodor, J. 1987: *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, J. and Pylyshyn, Z. 1988: Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, 28, 3–71.

Geis, F. and Levy, M. 1970: The Eye of the Beholder. In R. Christie and F. Geis (eds) *Studies in Machiavellianism*. New York: Academic Press.

Goldman, A.I. 1989: Interpretation Psychologized. *Mind and Language*, 4, 104–19.

Goldman, A.I. 1992: In Defense of the Simulation Theory. *Mind and Language*, 7, 104–19.

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. and Wellman, H.M. 1992: Why the Child's Theory of Mind Really *Is* a Theory. *Mind and Language*, 7, 145–71.

Gordon, R. 1986: Folk Psychology as Simulation. *Mind and Language*, 1, 158–71.

Grandy, R. 1973: Reference, Meaning, and Belief. *Journal of Philosophy*, 70, 439–452.

Harris, P. 1989: *Children and Emotion: The Development of Psychological Understanding*. Oxford: Basil Blackwell.

Heal, J. 1986: Replication and Functionalism. In J. Butterfield (ed.) *Language, Mind and Logic*. Cambridge University Press.

Heal, J. 1995: How to Think About Thinking. In *Mental Simulation*. Oxford: Blackwell, pp. 33–52.

Hebb, D. 1949: *The Organization of Behavior*. New York: Wiley and Sons.

Kosslyn, S.M., Chabris, C.F., Marsolek, C. and Koenig, O. 1992: Categorical Versus Coordinate Spatial Relations: Computational Analyses and Computer Simulations. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 562–77.

Leslie, A. 1988: The Necessity of Illusion: Perception and Thought in Infancy. In L. Weiskrantz (ed.) *Thought Without Language*. Oxford University Press, pp. 185–210.

Leslie, A.M. and Thaiss, L. 1992: Domain Specificity in Conceptual Development: Neuropsychological Evidence from Autism. *Cognition*, 43, 225–51.

Lewis, D. 1979: Attitudes *De Dicto* and *De Se*. *Philosophical Review*, 88, 513–43.

Macdonald, C. and Macdonald, G. (eds) 1995a: *Connectionism: Debates on Psychological Explanation*. Oxford: Blackwell.

Macdonald, C. and Macdonald, G. (eds) 1995b: *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford: Blackwell.

McCulloch, W. and Pitts, W. 1943: A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5, 115–33.

Mealey, L. 1995: The Sociobiology of Sociopathy: An Integrated Evolutionary Model. *Behavioral and Brain Sciences*, 18, 523–99.
Minsky, M. and Papert, S. 1969: *Perceptrons*. Cambridge, MA: MIT Press.
Mozer, M. 1991: *The Perception of Multiple Objects: A Connectionist Approach*. Cambridge, MA: MIT Press.
Nasby, W., Hayden, B. and dePaulo, B.M. 1979: Attributional Bias Among Boys to Interpret Unambiguous Social Stimuli as Displays of Hostility. *Journal of Abnormal Psychology*, 89, 459–68.
Nichols, S., Stich, S., Leslie, A. and Klein, D. 1996: Varieties of Off-Line Simulation. In P. Carruthers and P. Smith (eds), *Theories of Theories of Mind*. Cambridge University Press.
Perner, J., Leekam, S. and Wimmer, H. 1987: Three-year-olds' Difficulty with False Belief: The Case for a Conceptual Deficit. *British Journal of Developmental Psychology*, 5, 125–37.
Perry, J. 1979: The Problem of the Essential Indexical. *Noûs*, 13, 3–21.
Rosenblatt, F. 1962: *Principles of Neurodynamics*. New York: Spartan.
Rumelhart, D. and McClelland, J. 1986: *Parallel Distributed Processing*. Cambridge, MA: MIT Press.
Sejnowski, T.J. and Rosenberg, C.R. 1987: Parallel Networks that Learn to Pronounce English Text. *Complex Systems*, 1, 145–68.
Smolensky, P. 1988: On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences* 11, 1–74.
Smolensky, P. 1995: Constituent Structure and Explanation in an Integrated Connectionist/Symbolic Cognitive Architecture. In C. Macdonald and G. Macdonald (eds) *Connectionism: Debates on Psychological Explanation*. Oxford: Basil Blackwell, pp. 223–90.
Steinberg, M.S. and Dodge, K.A. 1983: Attributional Bias in Aggressive Adolescent Boys and Girls. *Journal of Social and Clinical Psychology*, 1, 312–21.
Stich, S. and Nichols, S. 1992: Folk Psychology: Simulation or Tacit Theory? *Mind and Language*, 7, 35–71.
Van Gelder, T. 1991: What Is the 'D' in 'PDP'?: A Survey of the Concept of Distribution. In W. Ramsey, S. Stich and D. Rumelhart (eds) *Philosophy and Connectionist Theory*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Von Eckhardt, B. 1993: *What is Cognitive Science*? Cambridge, MA: MIT Press.
Waas, G.A. 1988: Social Attributional Biases of Peer-rejected and Aggressive Children. *Child Development*, 59, 969–92.
Wellman, H. 1990: *The Child's Theory of Mind*. Cambridge, MA: MIT Press.
Wellman, H.M. and Woolley, J.D. 1990: From Simple Desires to Ordinary Beliefs: The Early Development of Everyday Psychology. *Cognition*, 35, 245–75.