Coherence as an Explanation for Theory of Mind Task Failure in Autism

DEEPTHI KAMAWAR, JAY L. GARFIELD AND JILL de VILLIERS

Abstract: O'Loughlin and Thagard (2000) present a specific computational implementation of the idea that the problems encountered by a child with autism in classic False Belief tasks derive from a failure to maintain coherence among multiple propositions. They argue that this failure can be explained as a structural feature of a connectionist network attempting to maintain coherence. The current paper criticizes this implementation because it falsely predicts that the same children will have a parallel problem with the False Photographs task. The fact that the content of representations makes a difference while the structure remains constant casts doubt upon their claim.

O'Loughlin and Thagard (2000) follow Frith (1970a, 1970b, 1989) and Frith and Happé (1994) in arguing for a 'weak coherence' explanation for the range of deficits observed in individuals with autism. On this theory this range of deficits is explained by the inability of the individual with autism to make use of coherence-based reasoning and hence of context in problem solving and in the interpretation of experience. This explanation stands in contrast to more familiar explanations of these phenomena in terms of a 'theory of mind' deficit (Baron-Cohen, 1991, among many others), or some combination of a specific theory of mind deficit, a linguistic deficit and/or a social perceptual deficit (de Villiers and de Villiers, 1999; Garfield, Peterson and Perry, in press). In this paper we take no position regarding the explanatory merits of the weak coherence proposal, *per se*, though we will quarrel with a recent computational version thereof.

O'Loughlin's and Thagard's contribution to this venture is the provision of a specific computational implementation of the weak coherence idea, an implementation they claim demonstrates that weak coherence so understood can provide an explanation of the failure of individuals with autism on standard theory of mind tasks such as unseen displacement problems (the Sally-Ann task).

1. The Computational Implementation of the Weak Coherence Thesis and our Quarrel with it

O'Loughlin and Thagard propose that the maintenance of coherence can best be understood as an instance of constraint satisfaction, and that this process is

Address for correspondence: Smith College, Northampton, MA 01063, USA.
Email: dkamawar@smith.edu; jgarfield@smith.edu; jdevilli@smith.edu

Mind & Language, Vol. 17 No. 3 June 2002, pp. 266–272.

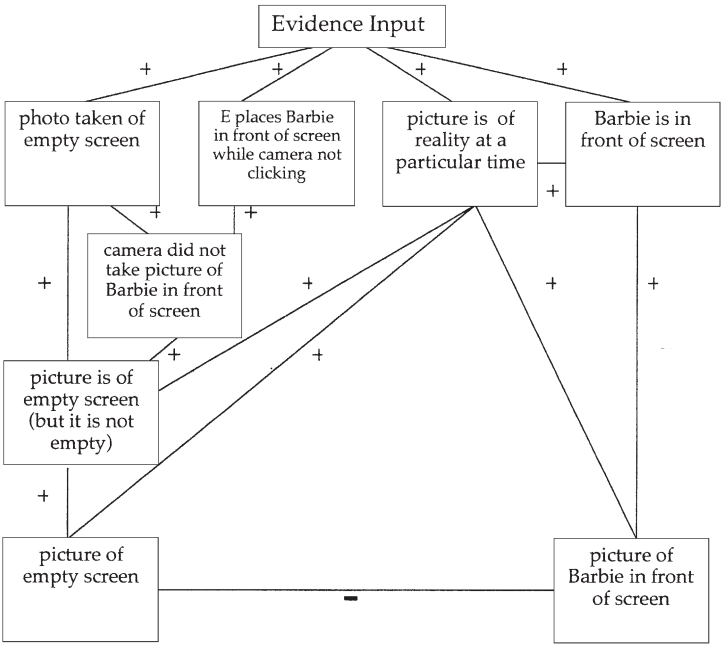

Figure 1 *O'Loughlin and Thagard's model of the Sally Ann task*

best represented in a connectionist network. They propose that failures to maintain coherence in such networks can best be explained by the network being trapped in a local maximum due to excessive inhibition on some or all of the links between nodes. They demonstrate, using several network examples, that increasing inhibition values along connections can cause the networks to be trapped in local maxima and that that this trapping can result in the failure of the system to maintain global coherence. Most significantly, such a failure can issue in the failure of the system to predict correctly the behaviour of the protagonist in the Sally-Ann task, mimicking the response of children with autism on this task. They conclude that this result explains the failure of individuals with autism on false belief tasks and so provides evidence both for the weak coherence theory of autism and for their connectionist computational model of weak coherence.

In Figure 1, O'Loughlin and Thagard model a standard theory of mind task (unexpected transfer task, Wimmer and Perner, 1983) in an effort to demonstrate that the failure of children with autism on such tasks is explained by a failure to maintain coherence. This network represents a portion of a child's knowledge used in predicting another's (Sally's) behaviour. The nodes containing the child's evidence and hypotheses are represented by the boxes while the lines connecting the boxes represent excitatory and inhibitory links (marked with '+' and '–' respectively).

The authors employed a computer program (ECHO) to compute coherence in this type of network. ECHO (using weightings given to it by the researchers) 'simulates the process by which people integrate the available information to reach the most coherent prediction about Sally's behaviour' (p. 382). Three different weighting combinations for excitation and inhibition were used. One (using a weight of 1.5 for false belief and a weight of –0.2 for inhibition) provided results consistent with children with autism's performance on false belief tasks. According to the authors, this demonstrates how a connectionist model could account for failure on the task by appeal to weak coherence and so can better account for problems on theory of mind tasks than can a '"theory of mind"/network structure explanation' (p. 384).

The *structure* of the network is what is meant to model weak coherence. It follows that this computational model can only explain these phenomena if the explanation is *structural* in character—that is, if it is the structure of the connectionist network, and not the *content* of the nodes that accounts for the network's performance. However, if the structure is doing the explaining, it would follow that any isomorphic network should demonstrate the same weak coherence, and it would follow from this that the performance of individuals with autism on isomorphic tasks should exhibit the same pattern of failure. As is well known, however, we do not observe this pattern. Children with autism pass 'false photo' tasks that are both *prima facie* isomorphic to false belief tasks and which can be plausibly represented by connectionist networks isomorphic to that presented by O'Loughlin and Thagard (Peterson and Siegal, 1998; Zaitchik, 1990) (see Figure 2).

2. The Form of the Argument

It is important to be clear about the form of O'Loughlin's and Thagard's argument. They offer a computational model of weakly coherent cognition, and that model is a connectionist network in which the state into which the network settles is a function of the relative strengths of inhibition and excitation between nodes. High inhibition values lead to the network settling in local maxima that fail to maintain coherence between interpretations of active nodes. The behaviour of a net that has settled into such a local maximum mimics the performance of subjects with autism on the unseen displacement task. This model is therefore taken to demonstrate that weak inhibition in a constraint satisfaction network is a plausible explanation of their failure on such tasks.

It is the fact that such a constraint satisfaction network is sensitive in this way to the adjustment of excitation and inhibition parameters that is supposed to explain the performances in question. Abandon this and the network model is superfluous. Therefore, if such a model is explanatory at all, then if individuals with autism solve problems of this kind using such a network, *any* problem

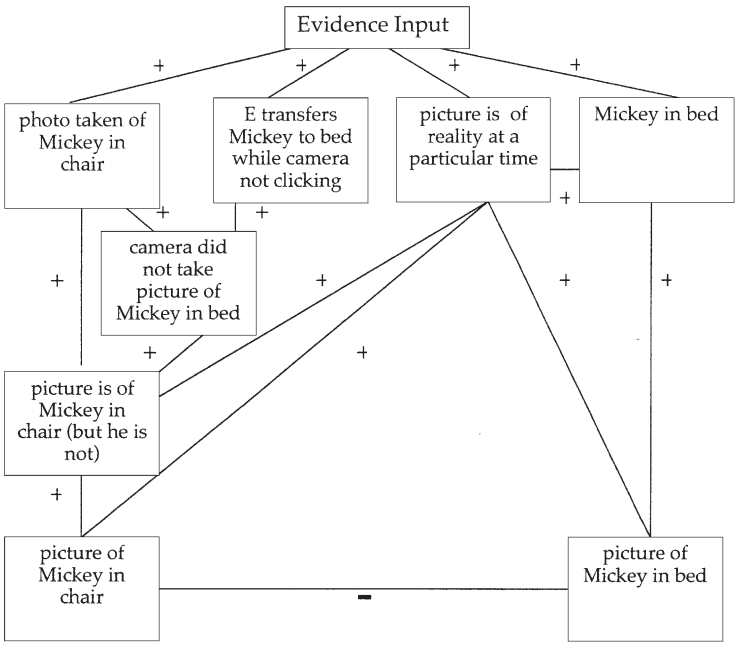

Figure 2 *A parallel model of the 'False Photo' task*

represented in a network isomorphic to the network in question should be solved in a similar way.

3. The Prediction of O'Loughlin's and Thagard's Model for the False Photo Task and the Falsity of that Prediction

The false photo task is isomorphic to the standard unseen displacement task. In each task an unobserved representation is constructed of a scene; the scene is changed in a way that ensures that the representation does not track those changes. In each case the experimental subject must then consider the new scene and determine whether the observable consequences of the representation will match the current scene or the original scene. The fact that some children with autism who fail the Sally-Ann task pass the false photo task has been taken by many as evidence that autism is not a general metarepresentational deficit, but instead a deficit specific to *mental* representations.

We note here that not only does a model of autism as a general metarepresentational deficit predict failure of individuals with autism on the false photo task, but that the O'Loughlin and Thagard weak coherence model predicts this failure as well. It makes this false prediction because it explains failure on Sally-Ann *not* by reference to the *content* of the nodes in question, but by reference to the *structure* of the network. There is nothing, as Figure 2 shows,

to distinguish the false photo from the false belief network *structurally*. If the latter is weakly coherent, so is the former. If weak coherence is the crucial feature of autistic thinking, thinking about false photos should be just as weakly coherent. And if so, subjects with autism should fail the false photo task. But they don't. We conclude that this explanation fails.

Of course one could try to patch things up by adjusting connection strengths in the false photo network so that inhibition levels are not so high, and so that the network is not misled by local maxima. But that would be terribly ad hoc. It would abandon the general weak coherence model of autism, and would then demand an explanation of why reasoning about the *mind* is weakly coherent, whereas reasoning about *other* representations is not, which brings us back to Theory of Mind deficit explanations.

Some (Russell, Saltmarsh and Hill, 1999) argue that the standard false photo task and the false belief task are not in fact isomorphic (i.e., their computational demands differ). Russell *et al.* contend that they differ in form because the standard false photo task does not make equivalent demands in *executive function* to the false belief tasks. While their argument may be relevant to some versions of the false-photo task it is beside the point in the present context, however, since the content of the standard false belief task is indeed perfectly matched by the change-of-location version of the standard false photo task we consider here (where the character moves from a bed to a chair). Note, however, that even if we were to grant the Russell *et al.* claim that there is a mismatch between the false belief and the false photo tasks because of their differential demands on executive function this would not undermine the argument we offer here against the explanatory force of O'Loughlin's and Thagard's model. For Russell *et al.* do not argue that these tasks differ in *structure,* but rather in respect of *content* (See Figure 3). Hence even if we grant that there is a difference, this difference is not a failure of isomorphism. According to the weak coherence model, the content of the representation should be irrelevant if the arguments are structured the same way. To the extent that the content ('nothing' versus another character in the picture) makes a difference in the modified false photo task, these results only count as further evidence *against* this explanation of failure on false belief tasks.

4. The Moral of the Failure of the Model

What do we learn from the failure of this model of autistic thinking? Just this: It is tempting to think that there is some very general feature of autistic thought that accounts for the entire, puzzling spectrum of deficits the syndrome comprises. This may be the case. It is also tempting to think that that general feature is a structural feature of thought. That *might* be the case, but there is no evidence that it is. One might be further tempted to think that that structural feature is weak coherence. We think that that idea needs to be made much more precise. To their credit, O'Loughlin and Thagard *do* make that

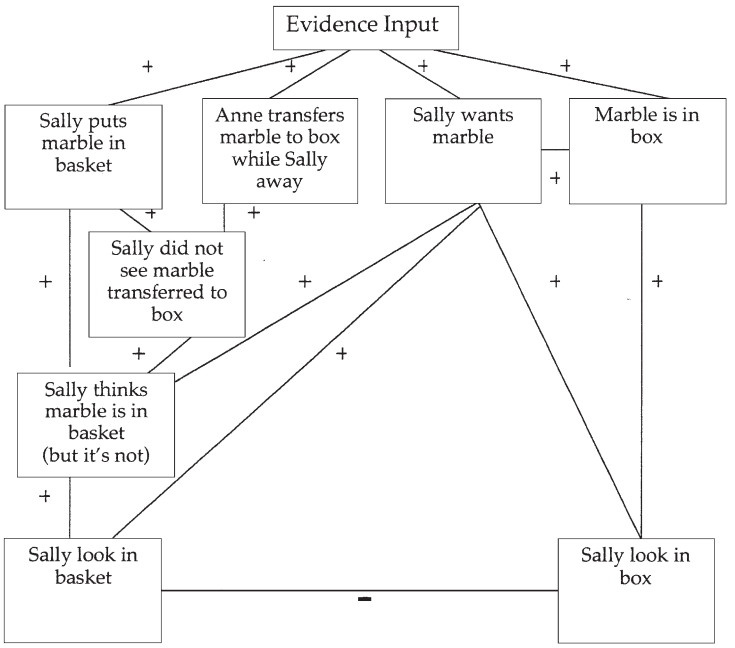

Figure 3 *A model of the revised False Photo Task in Russell et al.*, 1999

idea more precise. Unfortunately, once they do, it makes a false prediction. That is the advantage of precision. We think that whatever is the common core of autism, if indeed there is one, it is more specific to the mental than this model allows.

Departments of Psychology and Philosophy
Smith College

References

Baron-Cohen, S. 1991: The theory of mind deficit in autism: How specific is it? *British Journal of Developmental Psychology,* 9, 301–314.

de Villiers, J. and de Villiers, P.A. 1999: Linguistic determinism and the understanding of false beliefs. In P. Mitchell and K. Riggs (eds.), *Children's Reasoning and the Mind.* New York: Psychology Press.

Frith, U. 1970a: Studies in pattern detection in normal and autistic children I: Immediate recall of auditory sequences. *Journal of Abnormal Psychology,* 76, 413–430.

Frith, U. 1970b: Studies in pattern detection in normal and autistic children II: Repro-

duction and production of color sequences. *Journal of Experimental Child Psychology,* 10, 120–135.

Frith, U. 1979: *Autism: Explaining the Enigma.* Oxford: Basil Blackwell.

Frith, U. and Happé, F. 1994: Autism: Beyond theory of mind. *Cognition,* 50, 115–132.

Garfield, J., Peterson, C. and Perry, T. in press: Social cognition, language acquisition and the development of theory of mind. *Mind & Language.*

Leekam, S.R. and Perner, J. 1991: Do autistic children have a metarepresentational deficit? *Cognition,* 40, 203–218.

Leslie, A. and Thaiss, L. 1992: Domain specificity and conceptual development: Neuropsychological evidence from autism. *Cognition,* 43, 225–251.

O'Loughlin, C. and Thagard, P. 2000: Autism and coherence: A computational model. *Mind & Language,* 15, 375–392.

Peterson, C. and Siegal, M. 1998: Changing focus on the representational mind. *British Journal of Developmental Psychology,* 16, 301–320.

Russell, J., Saltmarsh, R. and Hill, E. 1999: What do executive factors contribute to the failure of false belief tasks by children with autism? *Journal of Child Psychology and Psychiatry,* 40(6), 859–868.

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

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