Developmental Science 1:2 pp 247–253

REPORT

Inhibitory processing in the false belief task: Two conjectures

Alan M. Leslie and Pamela Polizzi

Department of Psychology and Center for Cognitive Science, Rutgers University, USA

Abstract

Although it is well established that four-year-olds outperform three-year-olds on predicting behavior from false beliefs, this is only true when the false belief is coupled with a positive desire. Four-year-olds perform poorly in an otherwise standard false belief task when the protagonist's desire is to avoid rather than to approach a target. We account for this by assuming that the attribution of a false belief involves inhibitory processing. We present two versions of an inhibition model of successful belief-desire reasoning.

In a standard false belief task, a child is asked to predict the behavior of a protagonist who has acquired a false belief after an object is moved unexpectedly (Baron-Cohen, Leslie, & Frith, 1985; Wimmer & Perner, 1983). Around age four years, children first become able to solve such problems. No matter what accounts for this transition, a cognitive model of successful performance is required. We present two models of successful performance; both models give a key role to inhibitory processes.

Just as the usefulness of a currency depends upon a default attribution of genuineness, despite occasional forgery, so the usefulness of the concept, BELIEF, depends upon a default attribution of veracity, despite occasional falseness. Beliefs, like currencies, ought to be, and typically are, true. A sound strategy in belief attribution, then, is to attribute, by default, contents that are true. Accordingly, the attribution of a non-default belief – a belief with a false content – will require an extra processing step. We postulate that this extra step involves the *inhibition* of the default attribution-response. Only if the inhibition succeeds will the attributer search for an alternative content for the protagonist's belief. If the inhibition fails, the default belief attribution will go through and the protagonist will appear to the attributer to have a true belief.

Inhibition and false belief

Inhibitory processes are ubiquitous in psychological functioning. The center-surround organization of visual receptive fields (Hubel & Wiesel, 1968), mechanisms of shifting attention between targets (Posner & Presti, 1987), and executive planning of action (Shallice, 1972) are examples of inhibitory processes from very different levels of cognitive architecture. In regard to solving false belief problems, Carlson *et al*. (in press) show that decreasing inhibitory demands facilitates three year-old performance (see also Leslie & Thaiss, 1992; Hughes & Russell, 1993; Roth & Leslie, in press). In our models, the calculation of a false-belief involves first identifying a true-belief content, followed by the inhibition of that content. Inhibition allows attention to disengage from the true-belief content and move to the alternative, non-factual content of the false belief. The required inhibitory processing may depend upon the gradual development of prefrontal cortex (Goldman-Rakic, 1987; Bjorklund & Harnishfeger, 1990; Dempster, 1993). The successful four-year-old is capable of marshalling the required inhibition.

Refining the inhibition hypothesis

How can the inhibition hypothesis be tested? Assume a

Address for correspondence: Professor Alan M. Leslie, Center for Cognitive Science, Psych. Bldg., Busch Campus, Rutgers University, Piscataway, NJ 08855, USA. E-mail: aleslie@ruccs.rutgers.edu

task with two possible answers, A or B. A standard task demands a prediction of behavior from a belief-desire pair in which the belief is false and the desire positive. By hypothesis, recovering a false belief content intrinsically requires inhibition of a default. Whereas beliefs *ought* to be true, desires are not by default positive – the negative desire, *not to burn one's fingers*, is a perfectly ordinary desire. Still, a need for inhibitory processing in order to identify the target of desire can be created extrinsically. Suppose that the agent's desire is for whichever of targets A or B does *not* have property *x*. Under some circumstances, in order to identify which of A or B is the target of the agent's negative desire, one might first have to identify the target which *does* have *x*, say, A, in order then to identify the NOT(*x*) target of the desire, i.e., B. Having attended to target A, the brain must subsequently disengage from A and shift attention to B. Our models assume that the disengagement and shifting requires inhibition of attention to A. For our models, it is 'target shifting' rather than negativity which is critical. In fact, though a false belief can be entirely positive, our models say that attributing a false belief requires target shifting.

Suppose now that the subject is required to predict behavior from a false belief together with a 'target shift' desire. If we are correct, such a task will demand *double* inhibition. However, the two inhibitions cannot simply be summed to produce a stronger inhibition of the same target, because this will give the wrong answer (see below). Instead, the two inhibitions must interact so as to cancel each other out. We expect that *inhibiting an inhibition* will be hard, even though the four-year-old subject can comfortably marshal a single inhibition.

Testing the inhibition hypothesis

Cassidy (1995) modified a standard false belief task (with positive desire) to one in which the agent negatively desires the object. In this task, the agent wants to look in whichever container the object is *not*. Cassidy's four-year-olds all passed a standard false belief task. However, in the false belief with negative desire task, only 38% passed, a result usually associated with three-year-olds. For Cassidy, this result was entirely unexpected but is predictable from the double inhibition hypothesis outlined above. We therefore needed to see if this result would replicate.

A critical feature of our models is that in a false belief + negative desire task the two inhibitions will interact and not simply sum. Therefore, we reasoned that the difficulty of passing a true belief + negative desire task would not simply sum with the difficulty of passing a false belief + positive desire (i.e., standard) task to yield the difficulty of a false belief + negative desire task. Instead, there should be an interaction such that the difficulty of a task requiring double inhibition will be far greater than the sum of two tasks each requiring a single inhibition.

Our study can be thought of as a 2×2 design with factors *belief* (true, false) $\times$ *desire* (positive, negative), though we did not actually test with a true belief + positive desire condition because we assumed this would be trivially simple for four-year-olds. We made passing the standard false belief + positive desire condition an inclusion criterion to ensure that all subjects could marshal the required single (false belief) inhibition.

We also asked subjects in the false belief + negative desire condition a standard *Think* question. It is important to notice that double inhibition is *only* required by a *Prediction* question. Only in predicting behavior are belief and desire considered *together*: only then will the two inhibitions interact. We predicted that subjects' performance on the *Think* question would be significantly better than on the *Prediction* question.

Second, we wanted to see if a second inhibition could be introduced other than by way of a negative desire. For this purpose, we included a *Mixed-Up-Man* scenario. This involves a character who has positive desires but who always acts in a way 'opposite' to his desire. The only way to predict what the Mixed-Up-Man will do is first to identify what a normal man would do, inhibit that outcome, and choose the alternative.

Method

Subjects.

Ninety nine children were tested on a standard false belief task. Fifty seven of these passed and were tested further. Of these a further 8 were rejected for failing control questions. Included subjects were between the ages of 4:0 and 5:0 (mean age = 4 years 7.5 months) and were recruited from preschools and daycare centers in New Jersey.

Materials

Materials included three toy rooms constructed from foam board, one for each of the tasks (including screening task), distinctly colored boxes, and small dolls and props used to enact scenarios.

Procedure

We presented two tasks in story form, each with true and false belief conditions. Each subject was randomly assigned to two of the four conditions with the constraint that no child received both true and false belief versions of the same story. Sixteen children participated in each condition. A standard false-belief task modelled on the Sally and Ann task of Baron-Cohen, Leslie & Frith (1985) was administered either before or after the two tasks, counterbalanced across subjects. Children were required to pass the standard false-belief task for inclusion in the study.

Table 1 shows the task protocols.

Negative Desire task

A girl was described as not wanting to put food in a box containing a sick kitten, otherwise the kitten would eat the food and become worse. In the true belief condition, the girl watched the kitten move from box A to box B. In the false belief condition, she observed the kitten in box A but was absent when it moved to box B.

Opposite Behavior task

A 'mixed-up man' was described as always doing the opposite of what he desires. If an object is in box A, he would look for it in box B. In the true belief condition,

Table 1 *Protocols.*

Negative Desire task

This is Renee. Look!! She's got some food – it's a piece of fish. She wants to put the fish in a box. She is going to go inside to look for a box. [*goes inside, leaving fish behind*]

Here are two boxes. Let's look and see what is in them. In this box, there's a ball of wool. And in this other box, there's a ball of wool and there's also a poor, sick kitten. Renee does NOT want to give the poor little kitten the fish because it will make its tummy very sore. So she's going to go outside to get the piece of fish. She does NOT want to put the fish in with the sick kitten [*goes outside*]. Why does she not want to? Yes, *not* to make the poor kitten worse!

True Belief	False Belief
On her way back from getting the fish, look what Renee sees! The poor sick kitten crawls out of this box… and goes into this box. Did Renee see that? Yes!	Look what happens while she's gone! The poor sick kitten crawls out of this box… and goes into this box. Did Renee see that? No!

Look, now Renee has the fish.
Memory: In the beginning, where was the kitten?
Reality: Where is the kitten now?

Know: Does Renee know the kitten is in here? *Think*: Where does Renee think the kitten is?

Prediction: Which box will she go to with the fish?

Mixed-Up Man task

This is the 'Mixed-Up Man'. Do you know what he does? Every time he wants to do something, he does the *opposite*. If he wants an ice-cream, he eats a carrot! If he likes a cat, he pats a dog! If he wants something that is in here [box A], he looks in there [box B]. If he wants something in there [box B], he'll look in here [box A].

[*Man says*:] 'Look, there's a piece of candy in this box. I love candy, so I'll look in *this* (opposite) box for the candy.' [*take candy out of box*]. The Mixed-Up Man has a Mexican jumping bean. It jumps and wiggles around like this. Ok, one day, he puts his bean in this box. Then he goes on a walk. [*Exit.*]

True Belief	False Belief
On his way back, look what he sees! The bean wiggles and jumps into the other box! [*moves*].	While he's gone, look what happens! The bean wiggles and jumps into the other box! [*moves*].

Memory: In the beginning, where was the bean?
Reality: Where is the bean now?

Know: Does the man know his bean is in this box? *Think*: Where does the man think his bean is?

Prediction: Where is he going to look for his bean?

he watched as his Mexican jumping bean jumps from box A to box B, while in the false belief condition, he was absent as it moved.

Subjects who failed *Memory* or *Reality* (control) questions were excluded from further analysis. To maintain pragmatic naturalness, subjects were asked a *Know* question in true belief conditions and a *Think* question in false belief conditions. All subjects were asked a Prediction question.

In true belief conditions, passing requires indicating the location opposite to where the object is in reality. In the false belief conditions, passing requires indicating the box where the object actually is. Better performance was predicted in true belief than in false belief conditions.

Results

Table 2 shows the number of subjects by condition who failed *Control*, *Think* or *Know* questions. Excluding Table 2 subjects, Table 3 shows the number of subjects passing *Prediction* for each condition. Figure 1 shows more subjects passed *Prediction* for true belief than false belief conditions for both Negative Desire (Upton's $\chi^2 = 10.87$, $p < 0.001$, one-tailed) and Negative Behavior tasks (Upton's $\chi^2 = 9.02$, $p = 0.001$, one-tailed). Including subjects who failed the *Think* question (Table 1), *Think* (false belief) was much easier than *Prediction* (false belief) in both the Negative Desire (McNemar Binomial, $N = 11$, $x = 1$, $p = 0.006$) and Negative Behavior (McNemar Binomial, $N = 15$, $x = 0$, $p < 0.001$) tasks.

Table 2 *Subjects failing Memory or Reality Control questions or Think or Know questions*

	Condition	Memory/Reality	Think/Know
True belief	Negative Desire	1	1
	Negative Behavior	0	0
False belief	Negative Desire	0	2
	Negative Behavior	1	3

All subjects shown in this table were excluded from Table 2.

Table 3 *Subjects passing Prediction in each condition (n = 16 per condition)*

	Condition	
True belief	Negative Desire	15
	Negative Behavior	9
False belief	Negative Desire	6
	Negative Behavior	1

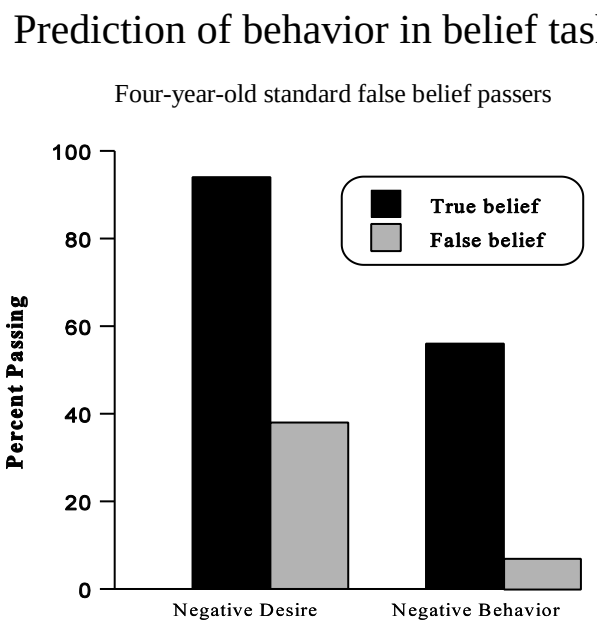

Figure 1 *Percentages of four-year-old children passing true and false belief tasks with negative desires and negative behaviors. All subjects pass standard false belief.*

Discussion

Our results confirm Cassidy's findings. Four-year-olds can predict behavior successfully from a false belief only when it is coupled with a non-target shifting positive desire.

Our model predicted that the difficulty of a double inhibition would exceed the sum of the difficulty of two single inhibitions. On the single inhibition deriving from a negative desire, only one of 16 subjects (6%) failed. On the single inhibition deriving from the Think (false belief) question, only 2 out of 18 subjects (11%) failed. If a double inhibition is simply the sum of two single inhibitions, then we should expect a 17% failure rate on the Prediction question in the false belief + negative desire condition. The observed failure rate was instead 62%, a result consistent with the prediction of our model.

It is not the case that just any information processing model would predict this result. We turned the positive desire of a standard false belief task into a negative desire by simply adding a 'not'. The burden for our subjects of this added 'not' was measured in the true belief + negative desire task and was minimal (6%

failure). However, when this same 'not' is added to a false belief task (which 86% of our subjects passed as measured by the Think question), the task is made dramatically harder. According to our model, the additional difficulty is not simply created by the word 'not' but by the specific consequences it has on problem processing. It is only if double inhibition is required that subjects who are above threshold for passing false belief will be pushed back down below.

The present results cannot definitively rule out alternative 'general difficulty' explanations. But we have three reasons for believing such alternatives are unlikely to be true. First, the 'Mixed-Up Man' scenario, notably, *without* negation, required a novel target shifting that introduced substantial general difficulty, as shown by performance in its true belief condition. This general difficulty had little impact on the calculation of false belief, as shown by only 3 failures out of 19 (16%) on Think (Table 2). However, performance on Prediction showed a similar 'excess' difficulty to that found in Negative Desire. Second, further studies in our laboratory of negation without target shifting suggest it does not produce 'excess' difficulty (Polizzi & Leslie, in preparation). Finally, Cassidy (1995) found that prediction from a false belief + negative desire was failed by 20% of adult subjects!

Two models of inhibitory processing in belief problems.

In belief tasks, the answers to the *where* test questions are 'targets', e.g., box A or B, that reflect the content of the attributed beliefs. According to our model, in tasks where there are essentially two possible targets for a character's belief, subjects first identify a target based upon an attributed true-belief, then inhibit that target, and select the alternative false-belief target: 'she thinks it's in *here* (B), no, in *there* (A).' If the subject is asked to predict the character's behavior, he must first make these belief-target calculations and then calculate the desire-target in relation to that belief-target. In the case of a positive desire, this latter calculation does not change the result so it adds little or nothing to the complexity of the calculation. Thus far, our model captures the finding that standard false belief tasks with Think questions are every bit as difficult as false belief tasks with Prediction questions.

Our model accounts for true belief tasks with negative desires in a similar way. First, the calculation of a true-belief-target is carried out and not inhibited. For a target-shifting negative desire, the only way to identify the target (in a two-alternative task) is first to identify the target X of the corresponding positive desire. Then, having identified the positive-desire-target, the subject in a negative desire task must inhibit that target and select the alternative, NOT (X).

Finally, when a prediction of behavior is called for in the case of a false belief + negative desire, the processing is necessarily more complex. We present two models between which we are unable to choose at this stage. In the first model (Figure 2), the target of true-belief and the target of positive-desire are identified in parallel. If the belief is false or the desire negative, inhibition is applied to this target and the alternative chosen, otherwise the original target stands. When prediction of behavior from false belief + negative desire is required, it is *not* sufficient to sum the inhibition by applying it twice. This gives the wrong answer. Instead, *inhibition*

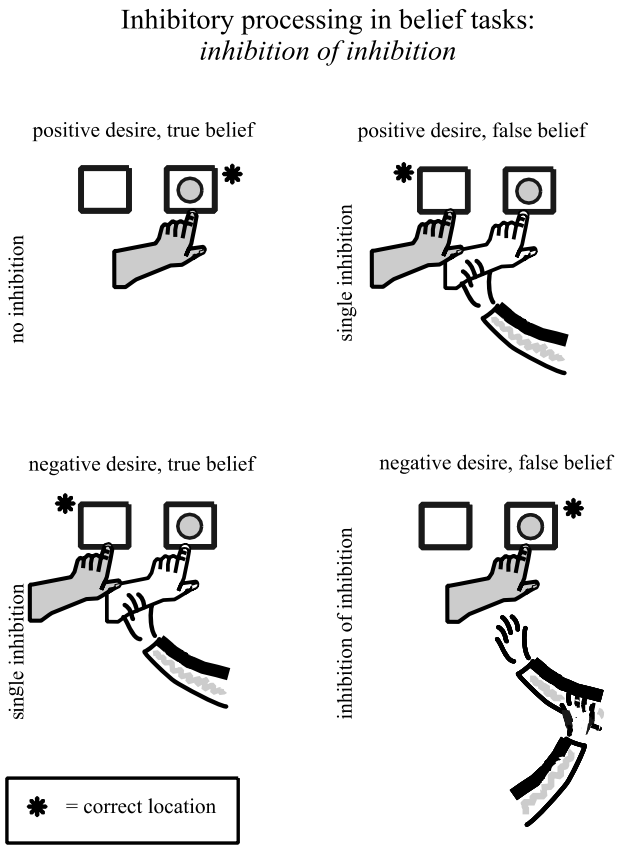

Figure 2 *Model 1: The target of the (true) belief and the (positive) desire are identified in parallel as shown in the top left panel by the shaded indexing hand. Subsequently, if either the belief is false or the desire negative, an inhibitory process is applied to this index, as represented in the next two panels as an inhibitory hand unshading the original index. Inhibition of this target leads to the selection of the other target. The final panel shows one inhibitory process inhibiting the other with the result that the original index is unchanged.*

of inhibition is required so that they cancel out. Since no inhibition reaches either the identified belief-target or the identified desire-target, the answer is exactly the same as for a trivially simple true belief + positive desire. However, marshaling a double inhibition to produce this answer strains executive processing.

In our alternative model, the belief target is identified first; the desire-target is then identified relative to this belief-target (Figure 3). In the doubled case, the subject is required to *return to a previously inhibited location*. After inhibiting the true-belief-target and selecting the false-belief-target, the subject identifies the positive-desire-target relative to the false-belief-target. Finally, this too must be inhibited. This forces the subject to revisit the target that was previously inhibited in the false belief calculation. As in other inhibitory contexts (e.g., Rafal & Henik, 1994), return to an inhibited target is difficult.

Both models show the inhibitory processes required for *correctly* solving belief tasks. Failure of any of the inhibitory processes will result in error. Specifically, failure to inhibit the true-belief-target will result in errors characteristic of three-year-olds.

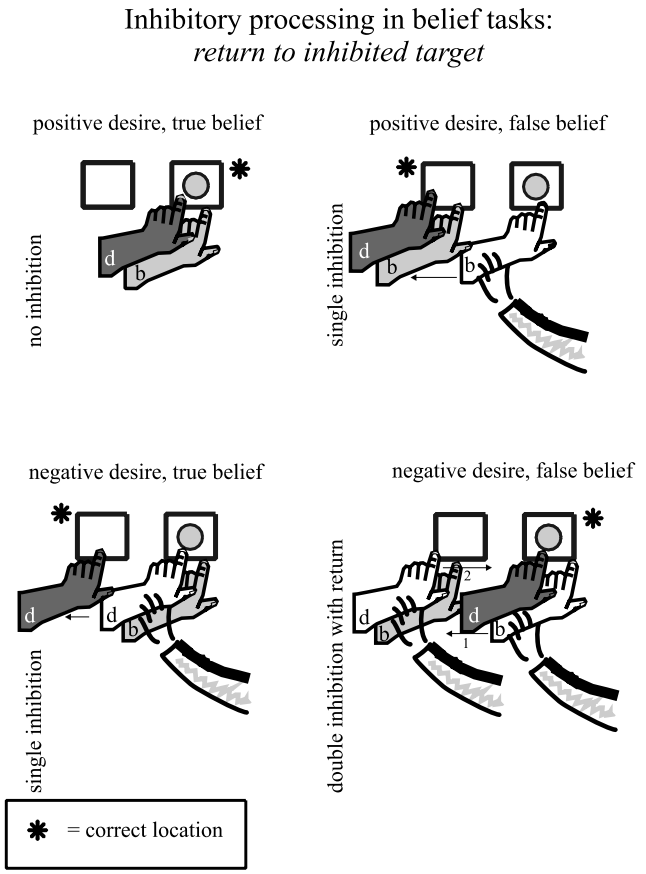

Figure 3 *Model 2: As shown in the top left panel, the target of true belief is identified first. The target of desire is then identified in relation to the belief target. Inhibition is applied, as appropriate, to the belief target (second panel) or to the desire target (third panel) causing the attribution target to shift. The final panel shows a sequence of target indentifications and inhibitions. First, the target of true belief is identified and inhibited causing the belief target to move to the alternative. Then the target of positive desire is identified in relation to the new (false) belief target. Finally, the positive desire target is inhibited in relation to the new (false) belief target. Finally, the positive desire target is inhibited forcing a return to the previously inhibited true belief target.*

Conclusion

We must stress that our four-year-old subjects did not fail to understand false belief. On the contrary, each and every one of our children who failed to predict the character's behavior had, only seconds before, correctly calculated that character's false belief. But when they entered that already made attribution into the calculation of the character's behavior, performance collapsed dramatically. Why that should happen is an important question for theories of the cognitive processes that solve false belief problems.

Acknowledgments

We thank Kimberly Cassidy for helpful discussions and two anonymous reviewers for comments on an earlier draft. We thank the following schools: Pixie Preschool, Children's House Montessori, Pilgrim Covenant Nursery School, Robbin's Nest, Good Day Nursery, and Bless U Child Day Care Center.

References

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

Bjorklund, D.F., & Harnishfeger, K.K. (1990) The resources construct in cognitive development: Diverse sources of evidence and a theory of inefficient inhibition. *Developmental Review*, **10**, 48–71.

Carlson, S.M., Moses, L.J., & Hix, H.R. (in press). The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development*.

Cassidy, K.W. (1995). Use of a desire heuristic in a theory of mind task. Paper presented to the *Biennial Meeting of the Society for Research in Child Development*, April 1995, Indianapolis, IN.

Dempster, F.N. (1992). The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. *Developmental Review*, **12**, 45–75.

Goldman-Rakic, P.S. (1987). Development of cortical circuitry and cognitive function. *Child Development*, **58**, 601–622.

Hubel, D.H., & Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology (London)*, **195**, 215–243.

Hughes, C., & Russell, J. (1993). Autistic children's difficulty with mental disengagement from an object: Its implications for theories of autism. *Developmental Psychology*, **29**, 498–510.

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

Posner, M.I., & Presti, D.E. (1987). Selective attention and cognitive control. *Trends in Neurosciences*, **10**, 13–17.

Rafal, R., & Henik, A. (1994). The neurology of inhibition: Integrating controlled and automatic processes. In D. Dagenbach and T.H. Carr (Eds.), *Inhibitory processes in attention, memory and language*. (pp. 1–51.) New York: Academic Press.

Roth, D., & Leslie, A.M. (in press). Solving belief problems: Toward a task analysis. *Cognition*.

Shallice, T. (1972). Dual functions of consciousness. *Psychological Review*, **79**, 383–393.

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

Received: 4 November 1997
Accepted: 5 March 1998