Psychological Science, in press

Mechanisms of belief-desire reasoning: Inhibition and bias

Ori Friedman

and

Alan M. Leslie

Department of Psychology and Center for Cognitive Science
Rutgers University

3607 words
Keywords: belief-desire reasoning; inhibition; selection processing; theory of mind

Acknowledgments: We thank Ellen Winner, Sara Baker, Bobbie Spellman, Angeline Lillard and three anonymous reviewers for a careful reading of and helpful comments on an earlier draft. This work was supported by NSF grant BCS-0079917 awarded to AML.

Corresponding author: Alan M. Leslie, Center for Cognitive Science, Rutgers University, 152 Frelinghuysen Road, Piscataway, NJ 08854
Tel.: 732 445 6152
email: aleslie@ruccs.rutgers.edu.

ABSTRACT

Biases in reasoning can provide insight into underlying processing mechanisms. We demonstrate a new bias in children's belief-desire reasoning. Six-year-old children were told a story in which a character is mistaken about which of three boxes contains some object. The character wants to go to one of the boxes, but only if it does not contain the object. In this scenario the character should avoid the box where she falsely believes the object to be, but might go to either of the remaining boxes. Though the character is equally likely to go to either box, children were biased to predict that the character would go to the box that contained the object. In a control task, the character had the same desire but did not have a false belief; in this case, children showed no bias, choosing each of the two correct answers equally. The observed pattern of bias was predicted by a developmental model of belief-desire reasoning. Competent belief-desire reasoning depends on a process of selection-by-inhibition in which the best belief content emerges from a set of candidates.

Competent social interaction depends on the ability to recognize and reason about people's mental states (what Premack and Woodruff (1978) termed 'theory of mind'). One cardinal component of this ability is the prediction of another person's action from that person's beliefs and desires. A key advance in investigating this ability was the development of the false belief (FB) task (Wimmer & Perner, 1983) and the discovery that belief-desire reasoning is well established before a child goes to school (Baron-Cohen, Leslie & Frith, 1985). In a standard version of the FB task, children are told about a girl, Sally, who sees a frog under a red box, but is absent when the frog moves to under a blue box. Following basic control questions, children must answer either a Think Question, about where Sally thinks the frog is, or an Action Question, about where Sally will look for the frog. Both questions have the same answer: Sally did not see the frog switch locations and so she mistakenly thinks that it is still under the red box, and will look for it there. We refer to the red box as the False Belief Location (FB-Location) because by indicating this location the child correctly attributes a false belief to Sally. We refer to the green box as the True Belief (TB-Location) because by indicating this location the child incorrectly attributes a true belief to Sally. A large number of studies consistently show that most children aged three years fail both questions (by indicating the blue box) but at four years most children pass (see Wellman, Cross & Watson, 2001 for review).

Most investigations of children's performance on the FB task have sought to identify the age at which children first attribute false belief, and until recently there has been little concern for discovering the processing mechanisms underlying successful performance. However, following the proposals of Russell, Mauthner, Sharpe, and Tidswell (1991) and of Leslie and Thaiss (1992) that success on the FB task might require inhibition, researchers have turned to investigate the relationship between inhibitory processing and belief-desire reasoning (e.g.

Carlson, Moses, & Breton, 2002; Perner, Lang, & Kloo, 2002; Roth & Leslie, 1998). In this paper we test between two detailed models of how inhibitory processing leads to success in FB tasks. Both models hold that success on FB tasks involves selecting between representations of alternative belief contents by inhibiting competitors.

Young children engage in belief-desire reasoning when their general knowledge and general reasoning powers are still severely limited. These contrasting abilities have led to the suggestion that the early capacity for belief-desire reasoning has an automatic, heuristic basis. Specifically, Leslie and colleagues have proposed that the young brain is equipped with a specialized 'theory of mind' mechanism (ToMM) that enables mental states to be attended to and learned about (Leslie, 1987, 2000). This proposed mechanism has a degree of neural specialization (Frith & Frith, 1999; Frith, Morton & Leslie, 1991; Gallagher & Frith, 2003; Gallagher, Jack, Roepstorff & Frith, 2002) and is present in children independently of general intellectual level (Leslie, 1987). Impairment of this mechanism may occur in certain complex genetic disorders, such as autism, with severe impact on social development (Baron-Cohen, 1995; Leslie, 1987; Leslie & Frith, 1990).

In the FB task, the role of ToMM is to ‘guess’ plausible and relevant beliefs that might be attributed to Sally. ToMM may make more than one guess, but one of these guesses should be that Sally believes what is *true*, (where ‘true’ reflects one’s own belief). This may be a useful heuristic because people’s beliefs about mundane things usually are true. But solving a FB task requires us to ignore our own ‘true’ belief. In order to ignore the ‘true’ belief, it must be inhibited, allowing the appropriate ‘false’ belief to be selected for attribution. The process of selection-by-inhibition is referred to as ‘selection processing’ (Leslie, 2000; Leslie & Thaiss, 1992; Roth & Leslie, 1998).

In a standard FB task, Sally wants to approach the object (for example, she wants to find the frog). In an avoidance desire FB task, Sally wants to avoid the frog and will, therefore, avoid the box in which she falsely believes it to be (FB-Location) and mistakenly go the true belief location (TB-Location) where the frog actually is. Cassidy (1998) found that an avoidance FB task is much more difficult than a standard FB task. Four-year-olds who passed the standard task failed the avoidance FB task, performing like three-year-olds. Leslie and Polizzi (1998) obtained very similar results and proposed two models of selection processing, either of which could account for the results. Table 1 summarizes the ages at which children typically pass and fail approach and avoidance FB tasks. We will now briefly describe the two models of selection processing and then describe how we tested between them.

Table 1 about here

Two models of belief-desire reasoning

The basic idea behind both models is that, in attributing beliefs to other people, an unconscious "theory of mind mechanism" (ToMM) automatically provides a small set of belief contents that might plausibly be attributed. Then, an attentional/decision process in the brain — selection processing — selects the most plausible content from this set. These ideas are reminiscent of certain approaches to the role of inhibition in visual attention (e.g., Klein, 1988; Posner & Cohen, 1984; Watson & Humphreys, 2000).

The selection process can be visualized as involving a mental pointer or index, which is attracted to the currently most salient content. Both models share the following assumptions:

1. The salience of contents varies by degree and contents can have their salience decreased by inhibition.

2. Among the contents, there is always one which is true, the 'true-belief content'. Initially the true belief content is most salient.

3. For success in a FB task, the true-belief content must be inhibited so that its salience falls below that of the false-belief content; otherwise a true-belief will be attributed.

4. Predicting the behavior of a character with an avoidance desire requires first identifying (indexing) the target to be avoided and then inhibiting that target so that it will be avoided.

In both models, predicting a character's behavior requires one inhibition in approach FB tasks, but two inhibitions in avoidance FB tasks. The models differ from one another in three main respects: first, in whether selection is done serially or in parallel; second, in how many selection indexes are used; and third, when predicting action in an avoidance FB task, in how the desire and belief inhibitions combine.

In model 1, "inhibition of inhibition", there is only a single index and inhibitions apply in parallel. When predicting where Sally will go in an avoidance FB task, the target of true belief (and approach desire) — the TB-Location — is indexed first. Because Sally's belief is false and her desire is to avoid the frog, two inhibitions are required, and are applied such that they cancel out by inhibiting each other. The end result is that the TB-Location remains indexed, and is correctly selected as the location where Sally will go. Panel A of Figure 1 depicts the operation of Model 1.

In Model 2, "inhibition of return," two indexes are used, one for belief and one for desire. Furthermore, inhibitions are applied serially, with belief inhibitions applied first and desire inhibitions second. To predict where Sally will go in an avoidance FB task, the target of true belief is again initially indexed. Because Sally's belief is false, the first of the serial inhibitions is applied, causing the belief index to move to the FB-Location. Next, the target of desire is identified, its index being attracted to the FB-Location (because it is more salient). However, because Sally's desire is to avoid the target, the desire index must be inhibited. For a successful prediction, the desire inhibition must be sufficient to lower the salience of the FB-Location below even that of the previously inhibited TB-Location. Panel B of Figure 1 depicts the operation of Model 2.

Figure 1 about here

Testing between the models

These models were developed to account for two-location tasks in which there is a single correct answer. Avoidance FB tasks, however, can be devised to have two (or more) equally correct answers by having three (or more) locations. If Sally believes (falsely) that the frog is in one box and she wants to avoid it, she might equally well go to either of two remaining locations, one of which contains the frog and the other of which does not. When we examined the models in the context of this three-location task, we found that they made opposite predictions about *which* correct answer would be preferred. We can therefore use such tasks to test the models. In a task with two equally correct answers, subjects might choose both answers equally often or they might show a systematic bias to one of the answers. Biases in reasoning can be used to probe underlying processing mechanisms (e.g., Tversky & Kahneman, 1974). If subjects choose each correct answer equally often, then both models will be refuted; if, however, subjects systematically select one of the answers, then one model will be refuted and the other supported. Figure 2 illustrates the action-prediction generated by each model in a three-location task.

Figure 2 about here

For model 1, the addition of a third Neutral-Location makes no difference because the canceling out of inhibitions leaves the TB-Location as the preferred answer. For model 2, however, the addition of a third Neutral-Location means that there is a correct answer that does not require the index to move to a previously inhibited location. Assuming that a target that has never been inhibited will be more salient than targets that have, the non-inhibited Neutral-Location should attract the desire index and will therefore be the preferred answer.

Experiment

The above predictions are made only for *successful* performance. Therefore, to test these predictions, we studied children who can pass avoidance FB. Pilot studies indicated that 6-year olds can have a high rate of success with avoidance FB. One group of children was given an avoidance FB task involving three locations (with two correct answers). Model 1 predicts that children who pass the 3-location task are more likely to answer the Action question by referring to the TB-Location than to the Neutral-Location. Model 2 predicts that passers in the 3-location task will prefer the Neutral-Location over the TB-Location. A third possibility is that children will show no bias and will either choose both of the correct locations or each of them half the time.

It is critical that we control for the falseness of the attributed belief. We therefore also used a task in which the character again wants to avoid the frog but this time her belief about the frog is true. However, there is a second frog which the character does not know about. We gave this avoidance Partial Knowledge (control) task to another group of children. In this control task, Sally sees a dirty frog under one box (TB-Location) but is absent when another dirty frog goes under a second box (Ignorance-Location). A third location (Neutral-Location) remains empty. In this task too the Action Question has two equally correct answers: Sally might go to the Ignorance Location because she doesn't know there is a frog there or to the empty Neutral-Location. Neither model predicts a bias in this task because there is no false belief to generate a belief inhibition, leaving action to be determined by the desire inhibition alone. Take model 1 for example: the desire inhibition is not cancelled out and so applies to the TB-Location, causing the index to move away to either of two equally salient alternatives (see Figure 3).

Figure 3 about here

Prior findings led us to predict a model 1 type bias in the avoidance FB task. A striking feature of the difficulty of avoidance FB tasks is that four-year-olds typically fail the Action question even though they pass the Think question immediately before it. For model 1, despite having already correctly calculated belief for the Think question, belief must be re-calculated along with desire in order to answer the Action question. Re-calculation is necessary because model 1 calculates belief and desire in parallel. For model 2, re-calculation is not necessary because belief and desire are calculated serially; belief can be calculated once in response to the Think question and then for the Action question desire identification is simply added to that result. Leslie, German, and Polizzi (submitted) reasoned that only if belief is re-calculated along with desire would it be possible to help children pass by using an 'easier' version of the Action question. An easier version of the Action question is to ask "Where will Sally look *first*?" — three-year-old children perform better with this question in the approach FB task (Siegal & Beattie, 1991; Surian & Leslie, 1999). Leslie, German, & Polizzi (submitted) found that in avoidance FB five-year-olds indeed performed much better with the 'look first' Action question, suggesting that, after a correct response to the Think question, they recalculated belief along with desire. This effect is predicted by model 1, not by model 2.

Method

Materials

Each task made use of a foam board stage, and dolls and props used to enact the tasks. All props were changed between tasks, except for the stage.

Procedure

Children who passed the screening task received either the avoidance FB task or the control task. Both tasks concern Sally, who wants to put a clean hat under a box but not with a dirty frog. In the avoidance FB task, Sally sees a frog under the middle box (FB-Location) of three boxes, but is absent when the frog moves to another location (TB-Location). In the control task Sally sees a frog under the middle box (TB-Location) of three boxes, but is absent when a second frog goes under one of the other boxes (Ignorance-Location). The Appendix contains the protocols for both tasks.

In both tasks, side of locations was counterbalanced across subjects, and children were asked the same Action Question, "Which box will Sally go to with her clean hat." This question has two equally correct answers.

We only examined responses to the Action question for children who first passed a series of Control Questions presented in fixed order. In the avoidance FB task the Control Questions were:

Memory: *In the beginning where did Sally see the dirty frog?*

Reality: *Where is the dirty frog now?*

Think: *Where does Sally think the dirty frog is?*

In the control task the Control Questions were:

Know 1: *Does Sally know that there is a dirty frog under the blue box?*

Know 2: *Does Sally know that there is a dirty frog under the yellow box?*

Subjects

Thirty-two out of forty-one children passed a standard FB task and were randomly assigned to receive either the avoidance FB task or the control task. To be included in the analysis, children were also required to pass all questions in these tasks, except for the Action Question. Two children in the avoidance FB task failed the Think question, and were rejected for that reason. The remaining 30 children were aged between 4:8 and 8:7 (mean = 6:7, SD = 1:5). In the avoidance FB task data was analyzed from 16 children aged between 4:9 and 8:7 (mean = 6:6, SD = 1:6) and in the control task from 14 children aged between 4:8 and 8:7 (mean = 6:7, SD = 1:6).

Results

Because our predictions concern only successful performance, of greatest interest are the responses of children who passed the Action Question. In the avoidance FB task, children were biased toward one of the correct answers, namely, the TB-Location. Two children chose the (incorrect) FB-Location. Of the 14 who passed, 13 children (93%) said the character would go to the TB-Location, and one child indicated both the TB-Location and the Neutral-Location, whom we conservatively scored as a Neutral responder (Binomial, N = 14, $x = 1$, $p < 0.002$, two-tailed). All children passed the control task, seven (50%) indicating the Neutral-Location, six (43%) choosing the Ignorance-Location, and one child indicating both the Neutral- and the Ignorance-Location, whom we conservatively scored as an Ignorance responder. In this task, there was no difference between the frequencies of children choosing either location (Binomial, N =14, $x = 6$, $p = 0.4$, n.s.). Table 2 shows responses across the two tasks. Responding across the two tasks differed significantly (Fisher's Exact, $p = 0.038$, one-tailed). Although children were free to do so, only 7% of children indicated both correct answers.

Table 2 about here

Discussion

We found a bias in children's belief-desire reasoning, using a task with two equally correct answers. In approach FB tasks, there is a single correct answer because the character has a desire to approach the target of false-belief. Tasks in which the character wishes to avoid the target of false-belief can have two (or more) equally correct answers. Though subjects might have chosen each location equally often, or responded with both correct answers, children who succeeded were instead biased to select the TB-Location. Responding could not have been spatial in nature because side of the TB-Location was counterbalanced across subjects. The bias was predicted by model 1, but was opposite to that predicted by model 2.

We included a control task to investigate alternative explanations for a model 1 bias. Suppose children simply prefer to select a location that currently contains an object over a correct but empty location ('object' bias). Or suppose children prefer to select a correct location that thwarts the character's desire ('failure' or 'irony' bias). Either of these biases, like model 1, favor the TB-Location. However, these other biases will also operate in the control task. The control task is very similar to the avoidance FB task in that it also has two equally correct answers to the Action question, namely, one location that contains a dirty frog and will defeat the protagonist's desire and one location that is empty. Again, the 'object' bias and the 'failure/irony' bias predict that children will favor the box containing a frog over the empty box. However, children in the control task showed no bias, picking each correct answer equally often.

In contrast to the 'object' and 'failure/irony' biases, the model 1 bias is sensitive to the character's epistemic state. Model 1, therefore, makes differing predictions for the two tasks: a TB-Location bias in the FB task and no bias in the control task (see Figure 3 for the model 1 mode of operation in the control task). To identify where Sally will go, first the target of true-belief and approach-desire — the TB-Location — is indexed. A desire inhibition (for avoidance) is then applied to the TB-Location causing the index to shift to either the Ignorance- or Neutral-Locations. Since neither of these locations has been inhibited, they should be selected with equal probability.

Using two similar tasks, we observed bias only in the avoidance FB task and not in the avoidance control task. The bias appears therefore to be related to the story character's epistemic state and not to extraneous biases. For children who succeed on avoidance FB tasks, attributing a false-belief to a character creates a preference for predicting one course of action over another, even though both are in fact equally likely. This finding is new, unexpected, and, whatever one's theoretical perspective, needs to be explained.

From our perspective, belief-desire reasoning is a process of selecting contents for beliefs and desires from among plausible alternatives. Selection of false-belief contents first requires inhibition of a default true-belief attribution. When desire attribution also requires inhibition, the two inhibitions interact, making the task difficult for younger children (Leslie & Polizzi, 1998; Leslie, German, and Polizzi, submitted). We have now shown that this interaction also leads to a selection bias in a 3-location task. These findings converge with those of Leslie, German and Polizzi (submitted) in supporting the 'inhibition of inhibition' model of selection. Competent reasoning about beliefs depends on the development of inhibitory control.

Appendix.

Protocol for the avoidance false belief and control tasks.

This is Sally and look what she has. It's a nice clean hat. Sally wants to put her clean hat away. So she puts it down outside and she goes into this room to look for a box to put it under. And look there are three boxes here. What color is this box? What color is this box? And what color is this box? Sally looks under the boxes. Is there anything under the red box? No, nothing. And under the blue box? Nothing. And under the yellow box? It's a frog and it's all dirty! Sally doesn't want to put her clean hat with the dirty frog because she doesn't want her hat to get all dirty.

Now Sally is going to go outside to get her clean hat, so she can put it under a box. Why does Sally not want to put the clean hat with the dirty frog? Right, because she doesn't want to get the hat all dirty. But look what happens while Sally is gone…

Avoidance FB Task	Control Task
The dirty frog hops from under the yellow box and goes under the blue box! Did Sally see that? No!	Another dirty frog comes in the room. And it goes under the blue box! Did Sally see that? No!
Memory: In the beginning where did Sally see the dirty frog?	*Know 1*: Does Sally know that there is a dirty frog under the blue box?
Reality: Where is the dirty frog now?	*Know 2*: Does Sally know that there is a dirty frog under the yellow box?
Think: Where does Sally think the dirty frog is?	
Action: Which box will Sally go to with her clean hat?	*Action*: Which box will Sally go to with her clean hat?

Table 1

Summary of previous findings on the development of belief-desire reasoning.

Check marks indicate typical pass performance, crosses typical failure.

	Age	
Task	3 ½ yrs	4 ½ yrs
avoidance True Belief	✓[a]	✓[a,b]
approach False Belief	✗‡	✓‡
avoidance False Belief	✗[c]	✗[a,b,c]

[a] Leslie, German, & Polizzi, submitted

[b] Leslie & Polizzi, 1998

[c] Cassidy, 1998

‡ standard finding in literature

Table 2

Location selected in answering the Action question in the avoidance false belief and control tasks. (Filled cells indicate incorrect answers.)

Avoidance False Belief Task		
TB	Neutral	FB
13 (81%)	1 (6%)	2 (13%)
Control Task		
Ignorance	Neutral	TB
8 (57%)	6 (43%)	0 (0%)

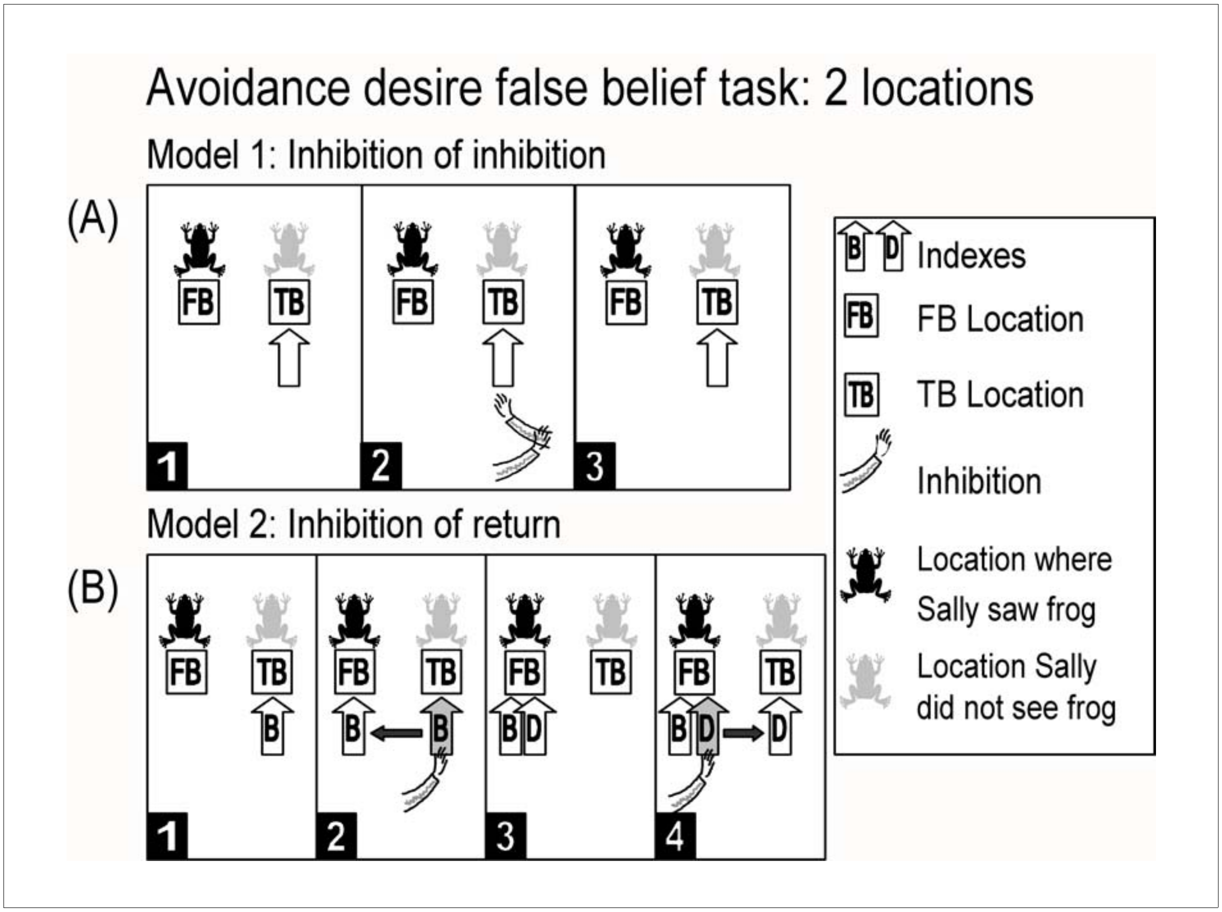

Figure 1. . Two competing models of selection by inhibition in an avoidance false belief task. In both models, the true-belief is initially more salient and attracts an index. In model 1, panel A, inhibitions for belief and desire cancel out, leaving the true-belief to be selected as the target of action. In model 2, panel B, the belief inhibition is applied to the true-belief causing the false-belief to become the more salient and to attract the index. A desire index is then applied to the false-belief location and subsequently inhibited, causing the index to move back to the true-belief location.

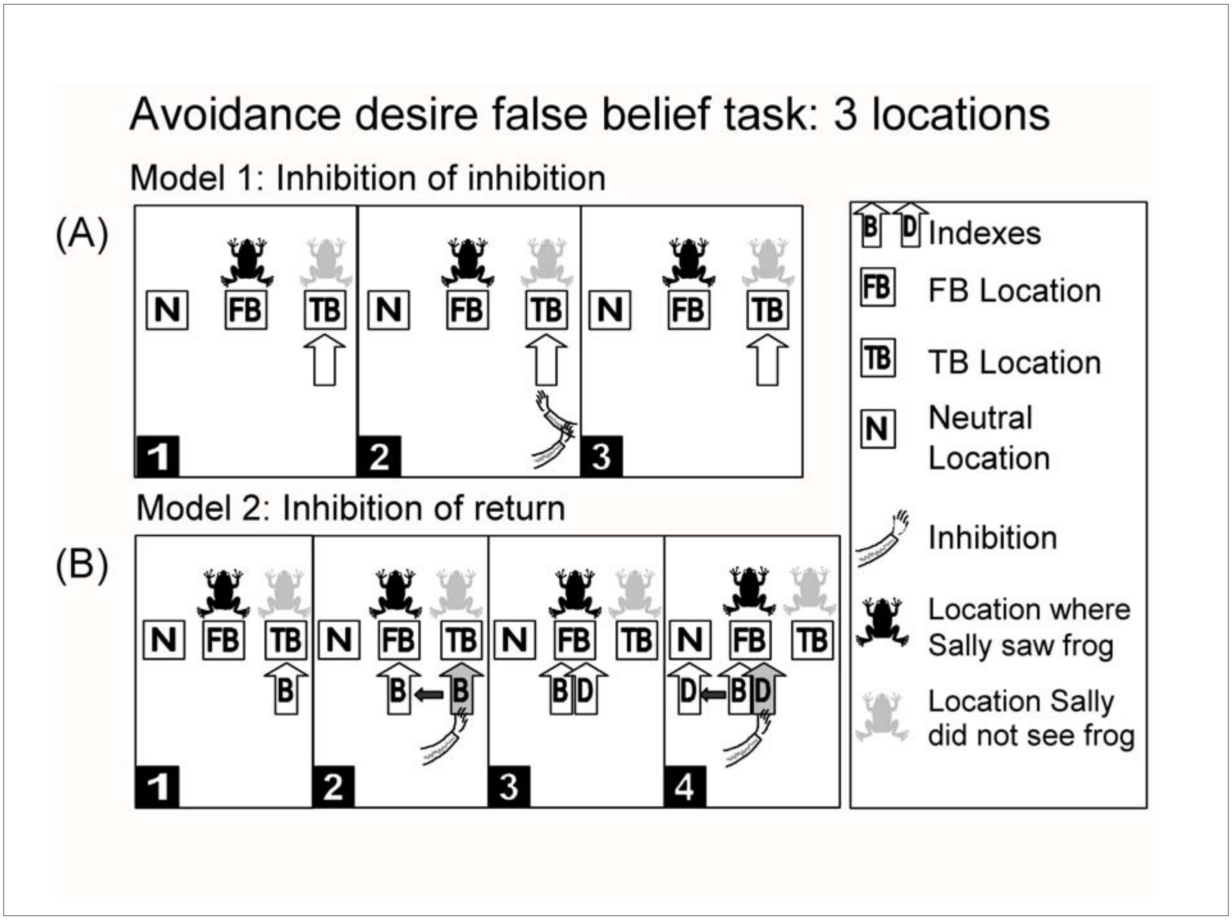

Figure 2. . Two competing models of selection by inhibition in 3-location avoidance false belief tasks. In model 1, panel A, the introduction of a third neutral location makes no difference to the outcome compared to a 2-location task, because the belief and desire inhibitions continue to cancel out, leaving the true-belief location selected. In model 2, panel B, the third neutral location will attract the index because in the end it is the only uninhibited location. A 3-location task provides a way to distinguish between models because each model predicts a different bias.

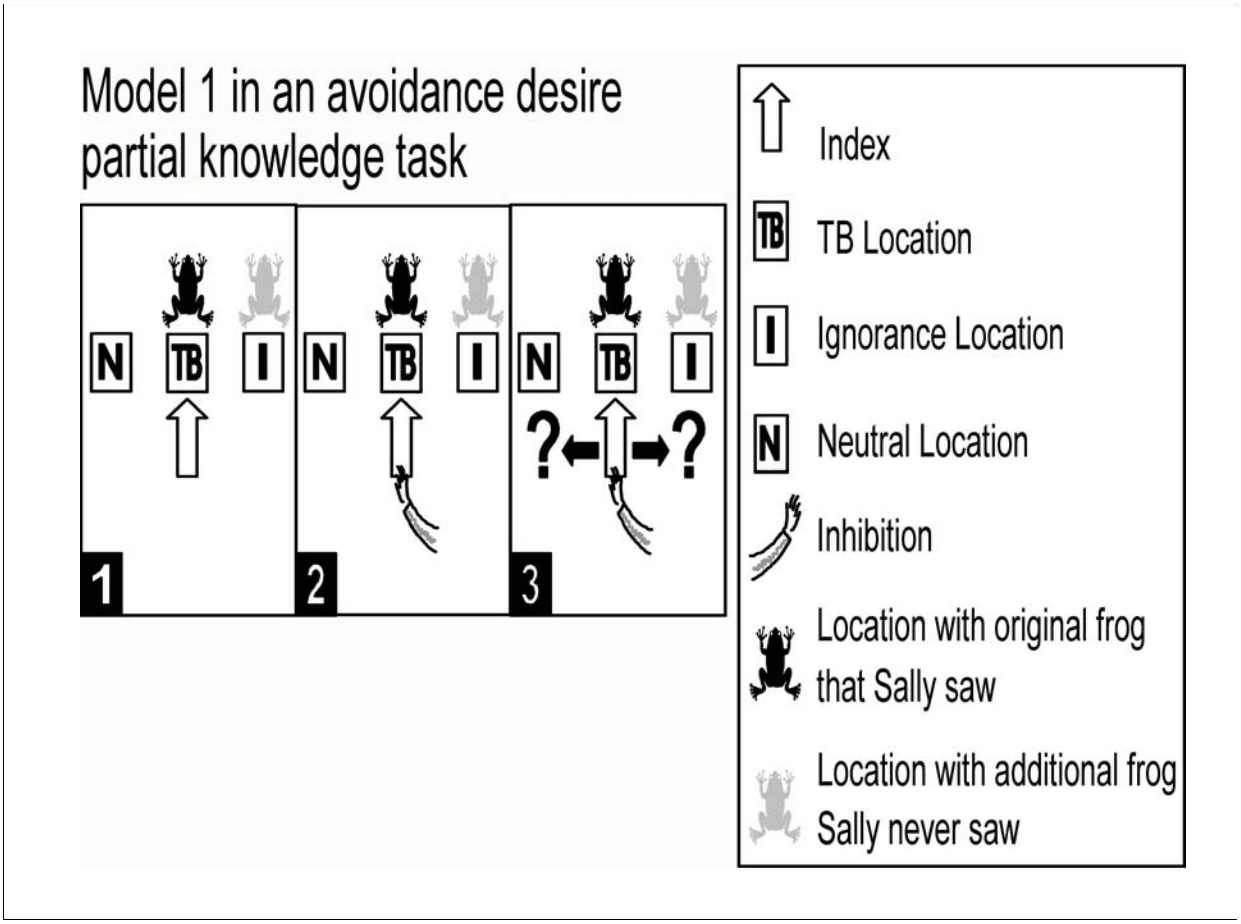

Figure 3. . When applied to an avoidance desire task in which a character knows about one frog under one box but does not know about another frog under another box, model 1 predicts no bias in selecting either of two equally correct answers.

References

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. MIT Press.

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

Carlson, S.M., Moses, L.J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant & Child Development*, **11**, 73–92.

Cassidy, K.W. (1998). Three- and four-year-old children's ability to use desire-and belief- based reasoning. *Cognition*, **66**, B1–B11.

Frith, C.D., & Frith, U. (1999). Interacting minds - A biological basis. *Science*, **286**, 1692–1695.

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

Gallagher, H.L., & Frith, C.D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences*, **7**, 77–83.

Gallagher, H.L., Jack, A.I., Roepstorff, A., & Frith, C.D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage*, **16**, 814–821.

Klein, R. (1988). Inhibitory tagging system facilitates visual search. *Nature*, **334**, 430–431.

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

Leslie, A.M. (2000). 'Theory of mind' as a mechanism of selective attention. In M. Gazzaniga (Ed.), *The New Cognitive Neurosciences,* 2nd Edition, (pp. 1235–1247). Cambridge, MA: MIT Press.

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

Leslie, A.M., German, T.P., & Polizzi, P. (submitted). Belief-desire reasoning as a process of selection.

Leslie, A.M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science*, **1**, 247–254.

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

Perner, J., Lang, B., & Kloo, D. (2002). Theory of mind and self-control: More than a common problem of inhibition. *Child Development*, **73**, 752–767.

Posner, M.I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D.G. Bouwhuis (Eds.), *Attention and Performance, X* (pp. 531–556). Hillsdale, NJ: Erlbaum.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral & Brain Sciences*, **4**, 515–526.

Roth, D., & Leslie, A.M. (1998). Solving belief problems: Toward a task analysis. *Cognition*, **66**, 1–31.

Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991). The `windows task' as a measure of strategic deception in preschoolers and autistic subjects. *British Journal of Developmental Psychology*, **9**, 331–349.

Siegal, M., & Beattie, K. (1991). Where to look first for children's knowledge of false beliefs. *Cognition*, **38**, 1–12.

Surian, L., & Leslie, A.M. (1999). Competence and performance in false belief understanding: A comparison of autistic and three-year-old children. *British Journal of Developmental*

Psychology, **17**, 141–155.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124–1131.

Watson, D.G., & Humphreys, G.W. (2000). Visual marking: Evidence for inhibition using a probe-dot detection paradigm. *Perception & Psychophysics*, **62**, 471–481.

Wellman, H.M., Cross, D., & Watson, J. (2001). Meta-analysis of theory mind development: The truth about false-belief. *Child Development*, **72**, 655–684.

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.