

Research Report

Is Belief Reasoning Automatic?

Ian A. Apperly,¹ Kevin J. Riggs,² Andrew Simpson,² Claudia Chiavarino,¹ and Dana Samson¹¹University of Birmingham, Birmingham, United Kingdom, and ²London Metropolitan University, London, United Kingdom

ABSTRACT—*Understanding the operating characteristics of theory of mind is essential for understanding how beliefs, desires, and other mental states are inferred, and for understanding the role such inferences could play in other cognitive processes. We present the first investigation of the automaticity of belief reasoning. In an incidental false-belief task, adult subjects responded more slowly to unexpected questions concerning another person's belief about an object's location than to questions concerning the object's real location. Results in other conditions showed that responses to belief questions were not necessarily slower than responses to reality questions, as subjects showed no difference in response times to belief and reality questions when they were instructed to track the person's beliefs about the object's location. The results suggest that adults do not ascribe beliefs to agents automatically.*

Reasoning about mental states such as beliefs, desires, and intentions is the stock-in-trade of people's everyday attempts to explain, predict, and manipulate human behavior. This ability—often termed *theory of mind*—is a fundamental component of human social cognition and of the uniquely human aptitude for communication (Baron-Cohen, Tager-Flusberg, & Cohen, 2001; Easton & Emery, 2005; Malle, Moses, & Baldwin, 2001; Repacholi & Slaughter, 2003; Sperber & Wilson, 1995). Surprisingly, however, there have been few direct investigations of the basic operating characteristics of theory of mind. For example, it is unknown whether inferences about beliefs, desires, and intentions are made automatically when people attend to the behavior of agents, or whether such inferences are made ad hoc, according to need. Knowing whether theory-of-mind inferences are made automatically is critical for understanding how theory of mind interacts with other activities such as communication, as well as for improving the paradigms available for investigating theory of mind with event-related methods of cognitive psychology or neuroscience.

Address correspondence to Ian Apperly, School of Psychology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK, e-mail: i.a.apperly@bham.ac.uk.

Several theorists have argued that theory-of-mind processes such as belief reasoning must be automatic (Friedman & Leslie, 2004; Sperber & Wilson, 2002; Stone, Baron-Cohen, & Knight, 1998). This argument draws upon evidence that belief reasoning may depend on cognitive processes that are domain-specific (Frith & Frith, 2003; Leslie & Thaiss, 1992; Saxe, Carey, & Kanwisher, 2004; though see Apperly, Samson, & Humphreys, 2005) or innate (Leslie, 2005; Onishi & Baillargeon, 2005). Domain-specificity and innateness are characteristic features of modular processes, and if theory of mind is modular, then it follows that processes such as belief reasoning may also be fast, informationally encapsulated, and automatic (Fodor, 1983, 2000; Friedman & Leslie, 2004; Leslie & Thaiss, 1992). The most detailed model of belief reasoning has been advanced by Leslie and his colleagues (Friedman & Leslie, 2004; Leslie, German, & Polizzi, 2005; Leslie & Thaiss, 1992), who have argued that belief inferences are performed by a fast, automatic, and domain-specific theory-of-mind module (ToMM) that parses the behavior of agents to generate a set of candidate belief contents. An executive-control process is responsible for the second step of selecting a single belief content from among these candidates. No direct evidence bears upon the automaticity of either of these processing steps.

In the current study, we used a novel incidental false-belief task to examine the automaticity of belief inferences. Our rationale was as follows. If subjects automatically parse events involving a human agent in terms of the agent's belief, then making a later explicit judgment about that belief will depend on information that has already been inferred and encoded. If this is the case, then judgments about belief might be made as quickly as judgments about other encoded information about the event. Moreover, explicitly telling subjects to keep track of an agent's belief should not result in faster judgments about that belief. In contrast, if subjects do not automatically infer and encode beliefs, judgments about an agent's belief, if subsequently requested, should be made relatively slowly compared with judgments about other information that *has* already been encoded. In this case, explicitly telling subjects to keep track of the agent's belief should result in faster judgments about that belief because subjects would have the opportunity to infer the belief in advance. We therefore compared the speed of subjects'

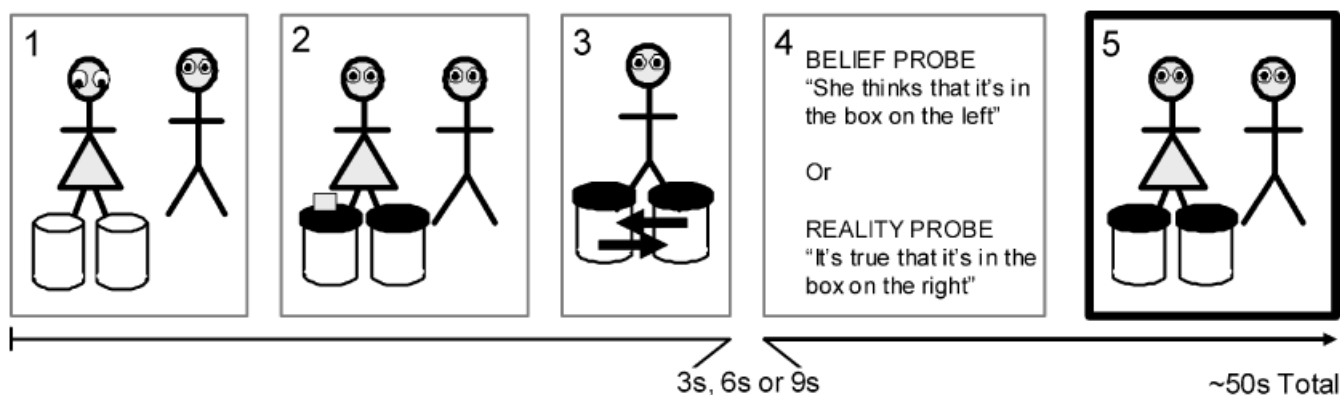


Fig. 1. Schematic event sequence for the experimental video trials. First, the woman looked in two open boxes (and gained true belief about the object's location). Second, the woman placed a marker to indicate the location of the object and then left the room. Third, the man switched the positions of the boxes (as shown here) or the location of the object, so that the woman had a false belief. Fourth, the woman returned (not shown here), and the probe sentence appeared (in all three conditions, response times were recorded from the onset of the probe sentences). Fifth, in Conditions 1 and 2 only, a change in the frame of the video then prompted the subject to point to the box containing the object.

responses to probe sentences about an agent's belief (where the agent thought an object was located) with the speed of responses to probe sentences about reality (where the object was really located). Critically, the speed of responses to the same probe sentences was compared across conditions in which we varied whether or not subjects were explicitly told to keep track of the agent's belief and the corresponding reality.

METHOD

According to the hypothesis that belief reasoning is automatic, subjects should infer beliefs even in the absence of any particular reason to do so. For Condition 1, we devised an incidental false-belief task in which subjects monitored relevant aspects of reality but had no particular reason to monitor agents' beliefs. We adapted video stimuli from a previous study (Apperly, Samson, Chiavarino, & Humphreys, 2004) so that probe sentences could be presented at unpredictable intervals to elicit belief or reality judgments from subjects. The exact event sequence varied across experimental and filler trials, but in all trials, a male actor hid an object in one of two boxes, and a female actor indicated where she thought it was hidden. Subjects had to identify the location of the object at the end of each trial. To do so, they needed to monitor movement of the boxes and take into account whether the woman had a true or false belief when she gave her clue.

Figure 1 depicts a generic event sequence for an experimental trial.¹ The subject saw the woman look in the boxes and then give a clue about the location of the object by placing a marker on one of them. By taking into account the fact that the woman's belief was true, the subject could infer the object's location. The wo-

man then left the room, and the man switched the locations of the boxes. This had two effects. First, the subject needed to update his or her representation of the location of the object in order to solve the task of locating the object at the end of the trial. Second, the switch resulted in the woman having a false belief about the object's location. The change in the woman's belief state was not relevant to the task of locating the object at the end of the trial. We were interested in whether subjects would, nonetheless, automatically infer the woman's new belief state. The video continued with the woman returning to the room. During this period, the video paused, and a probe sentence appeared: either a belief probe ("She thinks that it's in the box on the left [right]") or a reality probe ("It's true that it's in the box on the right [left]"). Probes were presented approximately 3 s, 6 s, or 9 s after the boxes were moved and the woman's belief became false. After the subject responded to the probe, the video continued until the appearance of a blue frame around the viewing area cued the subject to point to the location of the object. Because the only purpose of this component of the task was to encourage participants to keep track of information relevant for responding to reality probes, data from these pointing responses were not evaluated.

Given that subjects needed to maintain and update a representation of the object's location in order to point correctly at the end of the trial, we expected the reality probe to be answered with information that had already been processed (inferred and encoded). If subjects also maintained and updated a representation of the woman's belief, then belief probes would also be answered with information that had already been processed, and response times (RTs) to belief probes might be no different from RTs to reality probes. However, if subjects did not update their representation of the woman's belief automatically, then a correct response to the belief probe would require extra information processing, which might result in a slower RT for belief probes than for reality probes.

¹Half of the experimental trials followed a sequence that was similar except that rather than switching the boxes, the man transferred the object from one box to the other. Results did not differ for the two types of trials, which were therefore combined for all analyses.

Conditions 2 and 3 used the same video stimuli as Condition 1, and RTs were recorded for the same probe sentences. The key difference from Condition 1 was that subjects in Conditions 2 and 3 were explicitly instructed to keep track of where the woman thought the object was located. Thus, subjects were expected to infer the woman's belief in advance of the belief probe, meaning that any processing cost associated with this inference would not be reflected in RTs to belief probes. In Condition 2, subjects were explicitly instructed to keep track of both where the woman thought the object was located and where it was really located, and to point to the object's location at the end of each trial. In Condition 3, subjects were explicitly instructed only to keep track of where the woman thought the object was located. They were not asked to track where the object was located and were not required to point to the correct location of the object at the end of each trial.

Subjects

Undergraduate students participated for course credits or for a small honorarium. Twenty-four subjects (15 female, mean age = 21 years) were assigned to Condition 1, 24 (14 female, mean age = 20 years) were assigned to Condition 2, and 26 (17 female, mean age = 21 years) were assigned to Condition 3. Two subjects in Condition 3 failed to complete the experiment, and their data were not analyzed.

Design and Procedure

There were 24 experimental trials, 12 with a belief probe and 12 with a reality probe. The probe question for experimental trials was equally likely to refer to the box on the left and the box on the right, and the correct answer was always "yes." In addition, 56 filler trials were presented to reduce the likelihood that subjects would be able to anticipate the exact event sequence in the video or the timing, content, or correct answer for the probe sentences. Twenty-four trials used the same videos and probes as the experimental trials, but the correct answer was "no." Sixteen trials used the same videos as the experimental trials, but the probes concerned either physical facts other than the object's location or the knowledge state of the male actor. Sixteen trials combined all probe types with other videos using the same actors, objects, and events, but in different sequences (from Apperly et al., 2004).

Altogether, 80 trials (each approximately 50 s in length) were distributed over four experimental blocks, each comprising 6 experimental trials (3 with belief probes, 3 with reality probes) and a variety of filler trials. Within a block, correct answers were equally often "yes" and "no," and the object's location at the time of the probe was equally often on the left and on the right. Trials were presented in a pseudorandom order, avoiding consecutive experimental trials. The experiment was presented on a standard Pentium-based desktop computer using DMDX (Forster & Forster, 2003). RTs were recorded from the onset of probe sentences, and so reflected reading time for the probe sentence plus any other processing required for responding "yes" or "no."

RESULTS

For correct responses, RTs falling 2 standard deviations beyond the mean per subject per condition were removed. For belief probes, this criterion resulted in a loss of 10 data points (3.5%) in Condition 1, 11 data points (3.8%) in Condition 2, and 15 data points (5.2%) in Condition 3. For reality probes, this criterion resulted in a loss of 11 data points (3.8%) in Condition 1, 8 data points (2.7%) in Condition 2, and 10 data points (3.5%) in Condition 3.

Preliminary analyses showed no significant effect of the timing of the probes (3 s, 6 s, or 9 s after the woman's belief became false), and although subjects responded more quickly on earlier blocks than on later blocks, this effect was similar for belief and reality probes. Data were collapsed across these factors for further analysis.

An analysis of variance with probe type (belief, reality) as a within-subjects factor and condition as a between-subjects factor revealed a significant interaction between probe type and condition, $F(2, 69) = 3.49$, $p_{\text{rep}} = .93$, $\eta_p^2 = .092$. The main effect of probe type failed to reach significance, $F(1, 69) = 3.45$, $p_{\text{rep}} = .90$, $\eta_p^2 = .048$. The main effect of condition was non-significant, $F < 1$. Follow-up t tests showed that there was a significant difference between RTs to belief and reality probes in Condition 1 (incidental false-belief task), $t(23) = 3.51$, $p_{\text{rep}} = .99$ (longer RTs to belief probes than to reality probes), but not in Condition 2 (explicit belief and reality tracking), $t < 1$, or Condition 3 (explicit belief tracking), $t < 1$. The mean RTs are displayed in Figure 2.

In Condition 1, subjects made 35 errors in response to belief probes (12.2% errors) and 23 errors in response to reality probes (8% errors). This difference was not significant, but clearly indicates that the difference in RTs to belief and reality probes did not result from a trade-off between speed and accuracy. Subjects made 36 errors in response to belief probes (12.5% errors) and 33 errors in response to reality probes (11.5% errors) in Condition 2. The corresponding numbers in Condition 3 were

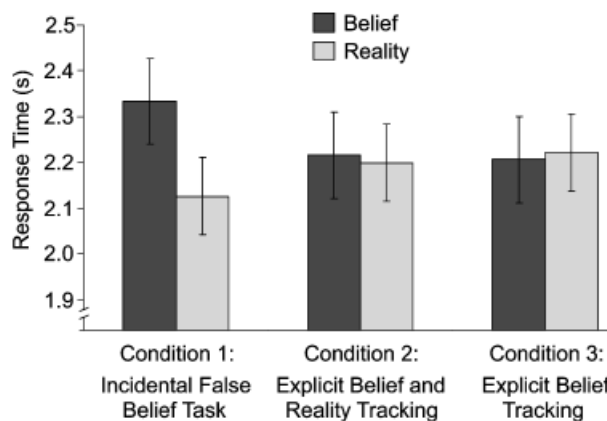


Fig. 2. Mean response times (bars represent standard errors) for belief and reality probes in Conditions 1, 2, and 3.

23 (8% errors) and 18 (6.3% errors). In neither of the latter two conditions was the difference between belief and reality probes significant.

DISCUSSION

The incidental false-belief task (Condition 1) showed a clear processing cost for belief probes in comparison with reality probes, a pattern consistent with subjects responding to reality probes using information they had already processed, but having to infer the woman's belief ad hoc in response to belief probes. This difference between probe types was absent in Conditions 2 and 3, suggesting that subjects could strategically infer the woman's belief in advance of the probes, and that after this inference was made, belief probes were not intrinsically slower to process or respond to than reality probes. The fact that responses to reality probes were as fast as responses to belief probes in Condition 3 (in which subjects were instructed only to keep track of the woman's belief) suggests that it may not have been possible to track the woman's false belief without also tracking the object's true location.

According to Leslie and his colleagues (Friedman & Leslie, 2004; Leslie & Thaiss, 1992), ToMM automatically parses the behavior of an agent to infer a set of candidate belief contents, but the process of belief ascription is complete only after a separate, executive selection processor selects the appropriate belief content from among these candidates. In these terms, responses to belief probes in the incidental false-belief task could have been relatively slow because ToMM had not inferred candidate belief contents, or because subjects' selection processor had not yet selected the appropriate belief content from the set provided automatically by ToMM. In the latter case, the current results would still be compatible with the involvement of an automatic subprocess (such as ToMM) in belief ascription. But in either interpretation, the process of ascribing a belief to the woman was incomplete by the time the probe appeared. Thus, whatever the nature of the subprocesses, the criterion for belief ascription—attributing a belief with a particular content to a particular individual—had not been met. In this most relevant sense, the current data suggest that belief reasoning is not automatic.

Our incidental false-belief task addresses a significant methodological problem in the theory-of-mind literature, the problem of knowing when a subject is making a theory-of-mind inference. The method identifies a narrow time window—immediately after the belief probe in the incidental false-belief task—within which belief ascription apparently takes place. Methods of this kind should enable event-related techniques from neuroscience and cognitive psychology to be used more effectively to investigate the nature of the complex component processes behind theory of mind.

Acknowledgments—This work was supported by a grant from the British Academy (SG-36063) to Ian Apperly.

REFERENCES

- Apperly, I.A., Samson, D., Chiavarino, C., & Humphreys, G.W. (2004). Frontal and left temporo-parietal contributions to theory of mind: Neuropsychological evidence from a false belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience*, *16*, 1773–1784.
- Apperly, I.A., Samson, D., & Humphreys, G.W. (2005). Domain specificity and theory of mind: Evaluating neuropsychological evidence. *Trends in Cognitive Sciences*, *9*, 572–577.
- Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D.J. (2001). *Understanding other minds: Perspectives from developmental cognitive neuroscience* (2nd ed.). New York: Oxford University Press.
- Easton, A., & Emery, N.J. (2005). *The cognitive neuroscience of social behaviour*. Hove, England: Psychology Press.
- Fodor, J.A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.
- Fodor, J.A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge, MA: MIT Press.
- Forster, K.L., & Forster, J.C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116–124.
- Friedman, O., & Leslie, A.M. (2004). Mechanisms of belief-desire reasoning: Inhibition and bias. *Psychological Science*, *15*, 547–552.
- Frith, U., & Frith, C.D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B*, *358*, 459–473.
- Leslie, A.M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences*, *9*, 459–462.
- Leslie, A.M., German, T.P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, *50*, 45–85.
- Leslie, A.M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition: International Journal of Cognitive Science*, *43*, 225–251.
- Malle, B.F., Moses, L.J., & Baldwin, D.A. (2001). *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: MIT Press.
- Onishi, K., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255–258.
- Repacholi, B., & Slaughter, V. (Eds.). (2003). *Individual differences in Theory of Mind: Implications for typical and atypical development*. New York: Psychology Press.
- Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*, 87–124.
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford, England: Blackwell.
- Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language*, *17*, 3–23.
- Stone, V.E., Baron-Cohen, S., & Knight, R.T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience*, *10*, 640–656.

(RECEIVED 10/20/05; REVISION ACCEPTED 3/14/06;
FINAL MATERIALS RECEIVED 4/18/06)