Cognition 111 (2009) 356–363

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Encoding of others' beliefs without overt instruction

Adam S. Cohen, Tamsin C. German *

Department of Psychology, University of California, Santa Barbara 93106-9660, United States

ARTICLE INFO

Article history:
Received 25 April 2008
Revised 14 March 2009
Accepted 15 March 2009

Keywords:
Theory of mind
Automaticity
Competence-performance
Modularity

ABSTRACT

Under what conditions do people automatically encode and track the mental states of others? A recent investigation showed that when subjects are instructed to track the location of an object but are not instructed to track a belief about that location in a non-verbal false-belief task, they respond more slowly to questions about an agent's belief, suggesting that belief information was not encoded or tracked automatically [Apperly, I. A., Riggs, K. J., Simpson, A., Samson, D., & Chiavarino, C. (2006). Is belief reasoning automatic? *Psychological Science, 17*, 841–844]. In the current experiments, we show that if belief probes occur closer in time to the events that signal the content of the agent's false belief, responses to those probes are faster than responses to probes about reality, and as fast as responses to probes about belief when instructed to track them, suggesting (i) beliefs may get encoded automatically in response to certain cues and (ii) that belief information rapidly decays unless it is maintained via 'top-down' instructions.

1. Introduction

The capacity to represent and reason about other people's minds, (sometimes called *mentalizing* or *theory of mind*), is a ubiquitous part of everyday social interaction. But under what conditions is this capacity deployed? Such thinking and reasoning might only occur when instructed and/or directed, such as when jurors are instructed to determine whether both *actus reus* and *mens rea* have been established beyond reasonable doubt (Model Penal Code § 2.01, 1962) or when friends ask each other about their thoughts on last night's baseball game. Alternatively, people might routinely and automatically track other people's beliefs and desires, as scenarios involving those people unfold, with no requirement for top-down instruction.

There are a variety of ways the term 'automaticity' is used (for review, see Bargh, 1989), and in the current paper, we use the term to refer to processing that is obligatory in the presence of appropriate input conditions (e.g. Fodor, 1983). For a given cognitive system, if specific cues are present and the system operates automatically, then processing is assumed to occur, independent of intention, endogenous attention, or awareness (termed "preconscious automaticity"; Bargh, 1989). The key question in considering automaticity for any cognitive process then will turn on identification of what the "appropriate input conditions" are for that system.

In the case of social cognitive reasoning, there is evidence that some social inferences, including mental state inferences, can occur without overt instruction (Uleman, Saribay, & Gonzales, 2008). In addition, if participants see "atypical" behaviors (e.g. an actor "pretending" to perform a simple action), there is increased activity in typical "mentalizing" areas, including medial prefrontal, inferior frontal, and temporoparietal areas (see e.g. Frith & Frith, 2006), even when subjects' explicit task is to monitor other aspects of the scenario, and there are no instructions to think about the mental states of the actors (German, Niehaus, Roarty, Giesbrecht, & Miller, 2004; see also Mar, Kelley, Heatherton, & Macrae, 2007). Similarly, passive 'violation of expectation' methods show that as early as 15 months of age, infants' expectations about an actor's search are sensitive to cues that define whether the actor's belief is true or false (e.g. whether the actor saw or did not see a change in the location of a target object; Onishi & Baillargeon, 2005).

* Corresponding author. Fax: +1 805 893 4303.
E-mail address: german@psych.ucsb.edu (T.C. German).

0010-0277/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2009.03.004

However, a recent empirical investigation has questioned the extent to which adult belief reasoning is automatic (Apperly, Riggs, Simpson, Samson, & Chiavarino, 2006). In a non-verbal, object displacement false-belief task, participants tracked an object's location over a short series of events (see Fig. 1, top row). Participants watched a character look into two physically identical containers, one of which contained the object, and then received a clue from the character to help locate the object. After receiving the clue, the agent left the room whereupon the object's location was switched, either visibly to participants (the object was removed from one container and placed in the other) or invisibly (the containers were swapped). At the end of each trial, participants were required to indicate the object's location (which they could infer from the placement of the actor's clue on invisible trials, and from the appearance of the object or the placement of actor's clue on visible trials) by pointing to one of the two containers.

On each trial, after the character returned, but before the participants were required to indicate the location of the object, a probe about reality (e.g. "It is true that the object is on the right") or belief (e.g. "She thinks that the object is on the right") assessed participants' encoding and tracking of these aspects of the event. Critically, participants *were not instructed to attend to the character's beliefs about the location of the object*; they were instructed only to track the object's actual location. Apperly, Riggs, Simpson, Samson, and Chiavarino (2006) argued that if belief reasoning is automatic, participants should attend to belief even without explicit direction. In that case, both reality and belief information would be readily available at the time of the probes: reality information because of overt instructions to track it, and belief information because of the "automatic tracking" of such information. Instead, Apperly et al. reported RTs to belief probes to be significantly slower than RTs to reality probes and concluded that belief reasoning is not automatic (Apperly et al., 2006, p. 844).

In the current paper we consider possible reasons for the limits on automaticity of belief calculation demonstrated by Apperly et al. (2006), and ask whether there might be conditions under which automatic encoding of belief happens in the kind of non-verbal false belief task used by those authors. An implicit yet critical feature of the Apperly et al. logic is that the responses to the belief and reality probes place equivalent performance demands aside from the content they assess. Notice, however, that while the response to the reality probe, which concerns the location of the object, is based on information updated at the time of the switch (either because the object is displaced or the containers swap; Fig. 1, top panel C), the response to the belief probe, which concerns the content of the belief (at the time of the switch, the truth value changes, but not the content of the belief), requires information that is acquired earlier in the proceedings (e.g. at the time when the actor places the clue; Fig. 1, top panel B). Therefore, the delay between the cues that lead to the acquisition of information relevant to the character's belief and the belief probe that assesses that information is extended compared to the delay between the cues that lead to the acquisition of information relevant to reality and the reality probes.

It is possible, then, that even though the belief information might have been encoded in response to those cues, the delay and intervening events ensure that the information is not available at the time of the probe. Interestingly, Apperly et al. (2006) varied the delay prior to the appearance of the probes (it appeared 3, 6 or 9 s after the object switch), and found no effect for either reality or belief. In the case of reality, because participants were actively instructed to track that information, it was likely maintained independent of the different delays. In the case of belief representations, if they were computed automatically, it

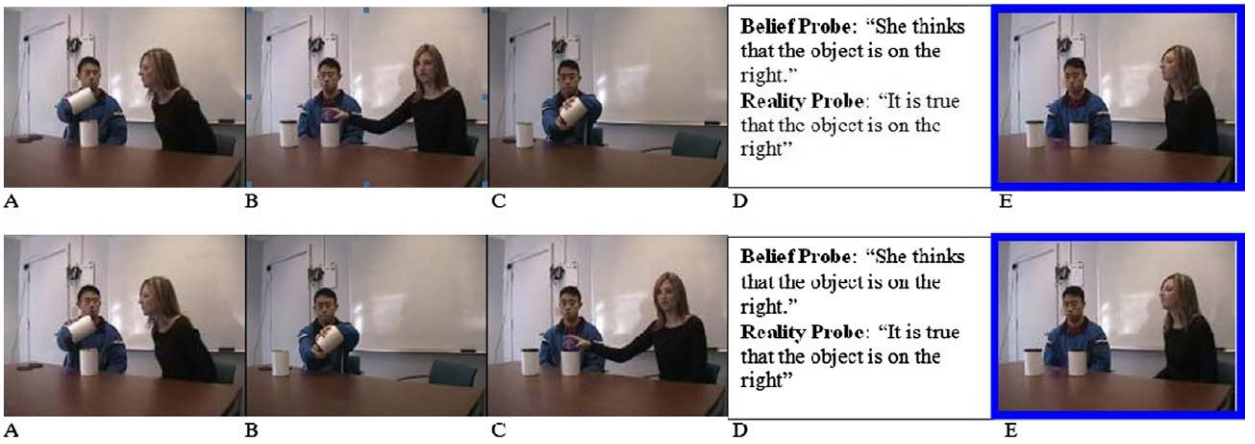

Fig. 1. The sequence of events in the two experimental conditions of Experiments 1 and 2. The top row shows the 'long delay' condition used in Apperly et al. (2006): (A) the woman looks into the containers, (B) gives her clue, (C) the man transfers the object to the other container (visible transfer trial shown), (D) the probe appears, and finally, (E) the blue frame appears indicating the end of the video. The bottom row shows the modified 'short delay' condition: (A) the woman looks into the containers, (B) the man transfers the object to the other container (visible transfer trial shown), (C) the woman gives her clue, (D) the probe appears, and finally, (E) the blue frame appears indicating the end of the video. Note the longer time interval in the long delay condition between the clue (B) and the probe (D) as compared to the shorter interval in the short delay condition where the clue (C) occurs later, and therefore closer to the probe (D). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is plausible that the encoded information had decayed to floor performance by the time the unexpected probe arrived. To draw an analogy with vision, perceptual information can surely be acquired under conditions where no "instruction" is given, but decays from visual short-term memory after just a few hundred milliseconds (Vandenbeld & Rensink, 2003). The lifetime of perceptual representations critically depend on mechanisms of attention (Rensink, 2000), and we would not want to conclude, simply as a result of them decaying without such maintenance, that the information was not initially acquired via an "automatic" process.

Are there conditions under which we might find evidence for the automatic encoding of belief in the non-verbal false belief task? In the experiments reported here, we test a set of circumstances in which the presence of information about a character's beliefs is probed after a shorter delay than that used by Apperly et al. (2006). A "short delay" condition for belief can be created by having the cues that provoke the encoding event (e.g. the agent's clue to what she believes) moved closer to the probe by having the character mark the container after the switch (see Fig. 1, bottom row). We can compare responses to belief probes in this short delay condition to the responses seen in Apperly et al.'s standard long delay condition.

If, as Apperly et al. (2006) argue, belief content is not encoded and maintained automatically, then the cost of calculating belief should exceed reality in both long and short delay conditions. However, if the clue to belief content results in encoding of that content even without explicit instruction, response times to belief probes will depend on the delay between the encoding of relevant cues and the appearance of the probes; they will be faster in the short delay condition than in the long delay conditions. Because of the overt instructions to attend to reality information, no difference in response times to reality probes between long and short delay conditions is expected.

An important point to note is that the switching of the order of events changes the nature of the belief attribution that might be made for some of the trials in the short delay condition, making them unsuitable for analysis. Critically, *for invisible trials*, where the container locations are switched (in contrast to visible trials, where the object is visibly moved from one container to the other), the first opportunity the participants have to learn the actual location of the object is when the actress provides the clue; however, in order to determine the object's location based on this clue, subjects must appreciate that the character has a false belief. The invisible, short-delay trials are thus unable to clearly assess automatic encoding of belief because encoding of belief is mandatory on these trials (in service of completing the explicit task). The visible trials, on the other hand, place no such demand on participants, who have already learned the location of the object when they saw it transferred between containers. On these trials, the information about belief is entirely irrelevant to the explicit task, and there is no reason for participants to encode it. These trials thus provide the key test of whether the cues to the character's belief might be sufficient to cause that belief to be encoded, despite no relevant task set.

2. Experiment 1

2.1. Method

2.1.1. Participants

Fifty undergraduates participated for class credit. Two were excluded owing to missing data. Of the remaining forty-eight, there were 29 females and 19 males (M = 19.75 years, SD = 2.178). Subjects were recruited through the University of California, Santa Barbara psychology department subject pool.

2.1.2. Design

A two (delay: long vs. short) × two (probe: belief vs. reality) repeated-measures design was used, with trials randomly presented across conditions. There were three test trials for each of the four conditions, creating 12 total test trials. Another 48 trials were filler videos randomly interspersed with the test trials to prevent subjects from predicting subsequent trial types. Each subject received 15 videos per block over four blocks, producing 60 total trials. Over the course of the experiment, correct answers to the probes were equally likely to be "yes" or "no" and the object was equally likely to be in the left or right container at the time of the probe. Videos ranged from 45 to 55 s in length depending on the particular video and were presented with E-Prime (Psychology Software Tools, Inc.) software. RTs were measured beginning at the onset of the probes.

2.1.3. Procedure

Prior to testing, subjects were instructed that they were going to see videos of a man and a woman. They were told that the man had placed an object in one of two containers and that the woman would help them find it. Their task was to point to the location of the object at the end of the video. Additionally, they were told that at some point during the video a statement would appear, and they would have to provide a "yes" or "no" button-press response to the statement. There were no instructions to track belief information.

Subjects saw two types of test videos: long delay, which were the same as Apperly et al. (2006), and the new short delay videos. In the long delay videos (see Fig. 1, top row), the man showed the woman the contents of the two cans. The woman then gave a clue about the location of the object by placing a marker on top of one of the two cans (encoding event). Next, she left the room, and while she was gone, the man changed the location of the object by moving the object from one can into the other. After the transfer, a probe interrupted the video. On test trials, one of two types of probes appeared: reality (e.g. "It is true that the object is in the can on the left") or belief (e.g. "She thinks that the object is in the can on the left"). On some filler trials, distracter probes (e.g. "It is true that the boxes have swapped"; "It is true that her shirt is black") were presented and on others the object was not transferred. After the subject responded to the probe with a yes–no button-press response, the video resumed and the woman returned to the room. When the video finished the screen froze and a blue frame appeared around the video to cue the subject to point to the location of the object.

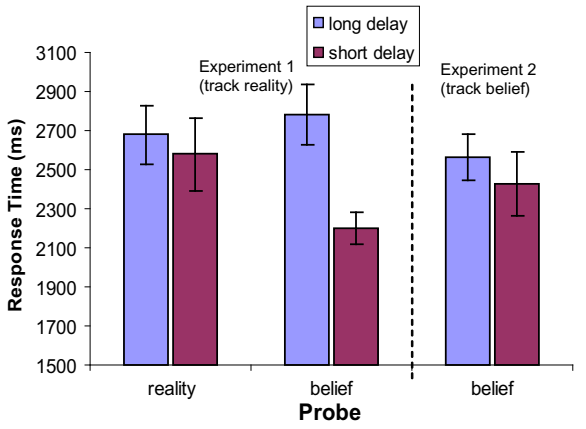

Fig. 2. The relationship between probe and delay on RTs in Experiments 1 and 2. While RTs are stable for reality probes across delays in Experiment 1, there is a significant difference for belief such that RTs for shorter delay is significantly faster than for long delay. There is also a significant difference between belief and reality in the short delay condition, such that RTs to belief are faster. RTs to belief probes under long and short delays are no longer different in Experiment 2 when belief tracking is overt. RTs to belief under short delay are no different in Experiment 1than Experiment 2 and RTs are not significantly different to belief probes across a long delay between Experiments 1 and 2 either.

In the short delay condition, the only difference was that the woman's clue (encoding event) came after the transfer (see Fig. 1, bottom row), immediately preceding the presentation of the probe. This generated a delay of about 3 s between when belief information could be encoded and the presentation of the probe (as compared to 23 s in the long delay condition).

2.2. Results and discussion

For reasons already discussed in the introduction, only visible transfer trials were entered in the analysis. RTs beyond three SD per subject were removed and not included in the analysis. This resulted in the following number of eliminated data points: four long-belief (3.2%), zero long-reality (0%), zero short-belief (0%), and zero short-reality (0%). Following Apperly et al. (2006), only correct "yes" responses were analyzed (although additional analysis revealed that the probability of a hit was not significantly different from the probability of a correct rejection).[1]

A two-factor analysis of variance (ANOVA) revealed a main effect of delay, $F(1,47) = 13.8$, $p = .0005$, $\eta^2 = 0.227$, where participants were significantly faster overall to respond in the short delay ($M = 2387$ ms, $SE = 119$ ms) as compared to long delay conditions ($M = 2731$ ms, $SE = 133$ ms). This main effect was qualified by an interaction between probe and delay, $F(1,47) = 4.74$, $p = .03$, $\eta^2 = 0.092$, depicted in the left panel of Fig. 2, that can be characterized as resulting from a much larger effect of delay on belief probes than on reality probes.

This characterization was examined with pairwise *t*-tests which revealed that indeed, RTs to belief probes were significantly slower under conditions of long delay ($M = 2783$ ms, $SE = 155$ ms) than they were under conditions of short delay ($M = 2197$ ms, $SE = 81$ ms; $t(47) = 3.74$, $p = .0005$). The same comparison for reality was not significant, $t(47) = 0.77$, $p = .44$. Additionally, in the short delay condition RTs for belief ($M = 2197$ ms, $SE = 81$ ms) were found to be significantly faster than RTs to reality ($M = 2577$ ms, $SE = 184$ ms), $t(47) = 2.45$, $p = .018$, $d = 0.35$. Finally, under long delay conditions, the original effect reported by Apperly et al. (2006) failed to be replicated, such that reality probes ($M = 2678$ ms, $SE = 153$ ms) were not significantly faster than belief probes ($M = 2783$ ms, $SE = 155$ ms), $t(47) = 0.67$, $p = .50$, $d = 0.09$.[2]

The results revealed that RTs to unexpected belief probes were faster in the short delay condition than they were under the long delay (where encoding of belief occurred earlier in the event sequence). On the other hand, RTs to reality probes were stable over the two delay conditions, suggesting that overt instructions to track reality removed the effect of delay.

The logic of Apperly et al. (2006) suggested that "automatic tracking of belief" would be demonstrated simply by equivalence between belief and reality probes. In fact, the observed result was that belief probes were responded to faster than reality probes, despite overt instructions to track reality, provided the probe appeared at a relatively short delay, suggesting that information about the belief of a social agent was in fact encoded in response to the behavioral cues emitted by that agent (i.e. the placing of a marker), *despite no explicit instruction to do so in task instructions*.

In the long delay condition, by contrast, encoded information about belief content, if it is indeed encoded, must also be maintained across a longer delay and several intervening events. With no instructions for participants to maintain that information, given that their task was to track the actual location of the object, we tentatively conclude that it decayed over time, such that reaction times were slowed in comparison to probes about the real location of the object. This analysis predicts that if subjects are instructed to track belief, there should be no significant difference in response times with respect to the delay between when the information was encoded and when it is probed for.

Apperly et al. (2006, pp. 841–842), argue that another potential sign of 'automaticity' in belief reasoning would be to observe no difference between responses under overt and covert conditions to track belief. Thus, a comparison between the response times to belief probes under the covert conditions (already observed in Experiment 1) and the response times to belief probes observed when subjects explicitly track belief information can shed light on this question.

[1] Errors for experiment 1 reported as percentages: long delay belief −5%, long delay reality −8%, short delay belief −9%, and short delay reality −12%, all *p*'s > .15, indicating that observed differences in RTs were unlikely due to a speed-accuracy tradeoff.

[2] One potential explanation for our failure to replicate Apperly et al. (2006) is that the two trial types, visible and invisible transfer, may make systematically different contributions to the condition means. One reason this might be the case is that when the transfer is invisible, greater demand is placed on maintaining the physical representation of the object (tracking the object is prioritized given the explicit task), which permits fewer resources to be allocated toward maintaining belief information, resulting in a response time cost. This suggests that collapsing across the two types of trials may not be warranted.

To test both of these ideas, in Experiment 2 we changed the instructions such that subjects were instructed to track the agent's belief about the location of the object through the unfolding events instead of tracking the object's actual location. This experiment was thus equivalent to Apperly et al.'s (2006) 'condition 3'. The reasoning behind this change was to allow a direct comparison between overt and covert belief reasoning while keeping the overall task demands the same; if the participants were required to perform the dual task of tracking belief and reality together (as in Apperly et al., 2006, condition 2), such a direct comparison would be potentially confounded by any change in overall task difficulty imposed by the addition of a second task.

3. Experiment 2

3.1. Method

3.1.1. Participants

Thirty-nine undergraduates (29 females and 10 males, M = 18.31 years, SD = 0.95) participated for class credit. Participants were different from those in the first experiment.

3.1.2. Design

Participants were assigned to both long delay and short delay conditions in a repeated-measures design, with trials randomly presented across conditions. There were three test trials for each of the four conditions, creating 12 total test trials. Another 48 trials were filler videos randomly interspersed with the test trials to prevent subjects from predicting subsequent trial types. The remaining features of the design were identical to that used for Experiment 1.

3.1.3. Procedure

The procedure was identical to that used in Experiment 2, except for instructions to track the woman's belief instead of the actual location of the object and being required to indicate, at the end of the video, where the woman believed the object was located instead of the actual location of the object, making conditions covert with respect to reality and overt with respect to belief.

3.2. Results and discussion

Data points falling beyond three SD were eliminated, including one in long-belief (1.0%), one in long-reality (1.0%), one in short-belief (1.0%), and zero in short-reality (0%). Only correct "yes" responses were analyzed. The results appear in the right panel of Fig. 2.[3,4]

The first hypothesis under test was that instructions to track belief should attenuate the effect of delay (long versus short) seen in Experiment 1. A repeated measures t-test revealed no differences between RTs to belief probes under long delay conditions (M = 2562 ms, SE = 119 ms) as compared to short delay conditions (M = 2427 ms, SE = 160 ms), $t(38) = 1.21$, $p = .23$.

The second hypothesis under test was to compare response times to belief under explicit instructions with response times under covert conditions. Under Apperly et al.'s (2006) logic, if belief reasoning is non-automatic, a cost should result when the belief probe occurs under covert conditions, as compared to when there are instructions to track belief. In fact, RTs to belief probes across short delay overt conditions in Experiment 2 (M = 2427 ms, SE = 160 ms) were not significantly different than RTs to the short delay covert condition of Experiment 1 (M = 2197 ms, SE = 81 ms), $t(85)$=1.35, p=.18. Note that the *direction* of the difference in means is actually the opposite to that required to provide evidence for the non-automaticity of belief attribution according to Apperly et al.'s argument. Similarly, there was no difference under a long delay between the overt belief condition in Experiment 2 (M = 2562 ms, SE = 119 ms) and covert belief in Experiment 1 (M = 2783 ms, SE = 155 ms) conditions, $t(85) = -1.08$, $p = .28$.[5]

The results of Experiment 2 support the claim that representations of belief encoded early or late in a series of events can be maintained under overt instructions to track that information. Combining these results with those from Experiment 1 we tentatively suggest that in the absence of such instruction, belief representations are subject to decay over time and/or interference from other cognitive processes.

4. General discussion

The findings presented here suggest that one of the factors contributing to the failure of participants to respond as rapidly to probes for belief information as they did to probes for information they were explicitly tracking in the Apperly et al. (2006) experiments might have been the fact that belief information, although encoded, was not maintained long enough to be available at the time that probes occurred. Here, in a modified version of the non-verbal belief reasoning task, the time over which the putatively encoded belief content was required to be held (Experiment 1) was shortened. The response pattern under such conditions reversed the effect observed by Apperly et al.: participants were *faster* to respond to belief probes than reality probes, *even when no instructions to track belief were given*. Importantly, this result was observed when the analysis was confined to trials for which it was entirely

[3] Errors for Experiment 2 reported as percentages: long delay belief −3%, long delay reality −1%, short delay belief −2%, and short delay reality −8%, indicating that observed differences in RTs were unlikely due to a speed-accuracy tradeoff.

[4] Mean response times and standard error for reality probes in Experiment 2 were M = 2670, SE = 117 for the long duration trials and M = 2662 and SE = 181 for short duration trials. Thus, based on these RTs, there was no reason to suppose that the overall task difficulty was harder when instructions were to track belief instead of reality (the mean response time for reality responses in the overt condition were 2678 and 2577 ms for long and short duration trials, respectively). Since these trials did not bear directly on predictions in either experiment, they were not analyzed further.

[5] One reviewer pointed out that, for belief under long delays, we might expect an RT advantage for overt compared to covert conditions given that RTs to belief in the covert case might be inflated owing to the fact that encoded information has decayed. We did not find this difference to be significant, but this is likely on account of the fact that performance in the long delay covert belief condition was not as poor relative to the other conditions as expected (see also footnote 2).

unnecessary to even calculate the agent's belief, because the visible transfer of the object provided sufficient information for the task to be completed.

In a second experiment, responses to belief probes under such "covert" conditions were shown to be no slower than responses to belief probes when subjects were explicitly instructed to track the agent's belief. According to Apperly et al. (2006, pp. 841–842), this pattern of responses would be a signature expected if belief reasoning involved automatic processing. Furthermore, the response time advantage observed under covert conditions for belief probes on short-delay trials compared to long delay trials disappears under overt conditions, supporting the explanation that mental state representations are automatically computed but readily decay under covert conditions.

It might be objected that the short delay conditions of Experiment 1 do not test "belief reasoning proper" because belief representations do not need to be maintained across events that result in them going out of date (e.g. the switching of the object's location), as happened in the scenario used by Apperly et al. (2006). As noted in the introduction, while the truth-value of the belief does go out of date, the probe used asks only about belief content, which does not change in either long or short delay conditions. Moreover, defining "belief reasoning" to include updating of truth values would be unprecedented: prediction and explanation of behavior based on mental state representations does not require us to know anything about *whether or not the content of beliefs are true or false*; it is sufficient simply to know what the content is.

These results leave open the possibility that at least some components of the mentalizing system may in fact be "automatic" in the sense that representations of belief content may be constructed without any explicit instruction to do so, and where there is no requirement imposed by the task set to do so, in response to certain cues present in streams of behavior. There is no reason at all in the short delay conditions with a visible transfer of object for participants to encode the character's belief, and yet responses to the subsequent belief probes are faster than are probes for information about reality, information that participants are instructed to track.

The large effect of manipulating 'delay' in this task suggests that representations created by potentially automatic sub-processes may have severe limitations in terms of their longevity and/or resistance to interference. The recruitment of additional systems (such as memory and attention) to maintain belief representations appears to be under "top-down" control.

This kind of division of labor is exactly that proposed in recent models of belief-desire reasoning, where a putatively automatic, "modular" subsystem generates candidate mental state representations that may or may not be selected for further processing by separable executive processes (e.g. German & Hehman, 2006; Leslie, Friedman, & German, 2004). The results reported in the current studies and in Apperly et al. (2006) are consistent with this kind of architecture for belief-desire reasoning, as is further evidence collected recently by Apperly, Back, Samson, and France (2008) who motivate their study by observing: "current studies of adults systematically confound the process of inferring a mental state with any processes involved in simply representing this information" (2008, p. 1094). Using a non-inferential false-belief task, these authors demonstrated processing costs to accrue from interference between belief and reality information. We endorse their emphasis on the need to understand the possibly wide range of executive processes recruited by belief-desire reasoning, many of which may also be "domain general" (i.e. also recruited across reasoning in other domains).

As a test of automaticity, at least as it has been articulated in some characterizations of the architecture of mentalizing (e.g. Leslie et al., 2004), the non-verbal false-belief task used in this study and in Apperly et al. (2006) has some limitations. First, the absolute duration of the RTs observed in Apperly et al. (2006), and in the studies reported here, is on the order of 2000–3000 ms. Uleman (1989) notes that single processes that are automatic typically require 300 ms or less. Minimally then, the belief reasoning as studied here likely reflects multiple processes. Two obvious processes, neither directly related to belief reasoning, each inflating RTs in this paradigm, include reading probe sentences and executing motor responses. It is questionable that the relatively long absolute RTs seen here can be used reliably to diagnose the presence (or not) of automatic processes.

Second, assessing the automaticity of "theory of mind" with offline measures in both of the current experiments as well as in Apperly et al. (2006) introduces a delay between belief processing and the response, and in doing so, disproportionate performance demands might potentially mask evidence of automatic processing, when the delays are not equated. When the probes occur appreciably later than the likely inputs that might provoke mental state inferences (e.g. language, communicative gestures, displays of emotion, or eye contact between the actor and actresses, etc.) the task inevitably measures the extent to which any encoded belief information might have been *maintained in the cognitive system*, rather than whether or not it was ever encoded.

Online assessments of belief-desire reasoning in real time, including behavioral measures such as eye-tracking or brain imaging techniques, that can exploit neural signatures associated with the processing of mental state content (e.g. fMRI, German et al., 2004; Saxe, Schultz, & Jiang, 2006; ERP, Liu, Sabbagh, Gehring, & Wellman, 2004), are methods that might help circumvent this problem by assessing the neural signatures time locked to the processing under investigation.

A final issue is that with which we opened this paper: exactly how should automatic processes, in the context of belief reasoning, be characterized? Proposals articulating a role for automatic sub-processes within mentalizing (e.g. Leslie et al., 2004), have argued that automatic processes do not operate "in a vacuum". Specifically, the parts of the mentalizing architecture argued to be "automatic" are assumed to produce representations *in response to certain cues*. That is, candidate representations are generated when certain inputs are presented to the system. These inputs might be 'behavioral' (e.g. 'goal directed action'; Wertz & German, 2007; Woodward, 1998; 'contingent interaction'; Johnson, 2003), derived from morphological

cues (e.g. 'directional eye gaze'; Calder et al., 2002; Masconi, Mack, McCarthy, & Pelphrey, 2005), or conveyed linguistically (e.g. utterances about mental states, Roth & Leslie, 1991; sentences labeling belief content, Apperly et al., 2008). At issue, then, is the availability and richness of social cues and their detection by systems that handle mental state inference. We propose, as has recently been argued for the case of visual attention (e.g. Kingstone, Smilek, Ristic, Friesen, & Eastwood, 2003), that researchers might learn much about the automaticity of mental state processing from the study of mentalizing in more ecologically valid contexts.

In the light of this argument, we consider the two recent imaging studies mentioned above that speak to the question of automaticity. First, using a task that could either be construed mentalistically or non-mentalistically depending on the task instructions, Saxe, Schulz, and Jiang (2006) observed a BOLD response in areas typically associated with theory of mind [viz. right temporoparietal junction (TPJ), and, less strongly, left TPJ] only when participants construed the scenario mentalistically. When reasoning about the same scenarios via a non mentalistic algorithm, there was no such response, a result the authors suggest speak against the idea that engagement of theory of mind regions happens automatically (Saxe et al., 2006, p. 295).

Second, using ERP, Liu et al. (2004) had participants watch a series of cartoon animations and then make a judgment about where a character thought an object was located or where the object was actually located. They reported an ERP component for "theory of mind" that separates from that for reality at about 800 ms post-stimulus and argued this peak was too late in the processing stream to signal an automatic, 'perception like' process (Liu et al., 2004, p. 995).

While these studies capitalize on the advantages offered by online measurement tools, note that the ERP study of Liu et al. (2004) actually locks the measured ERP signal to the presentation of a picture[6] presented subsequent to the events from which the belief calculation must be made. That is, the study measures the online neural signal associated with an instructed belief judgment based on events that occurred seconds before, rather than measuring the possible online belief processing associated with the unfolding event itself.

In the case of Saxe et al. (2006), this limitation is overcome because the BOLD response was measured across the animated trials, either while subjects responded based on mentalizing or via the response algorithm. Why then, in contrast to the results of German et al. (2004), was no neural signature of mental state reasoning seen when participants were performing an algorithm version of the task, but in the presence of content that might also be construed mentalistically? Our speculation relates to the possible extent of and/or richness of the cues that are the presumed to be the input to mental state reasoning mechanisms. German et al. (2004) presented real time video displays in which actors performed pretend or real actions, while performing a task unrelated to determining anything about the agent's action (judging whether the video clip was prematurely edited or not), and observed activations in typical mental state reasoning areas. The cartoon animations used by Saxe et al. (2006) were considerably impoverished stimuli by comparison, and may have provided insufficient input to invoke a mentalistic interpretation without top-down instruction. Consistent with this interpretation, Mar, Kelley, Heatherton, and Macrae (2007) show that activation in ToM areas (e.g. TPJ) and those implicated in the perception and interpretation of agency (e.g. superior temporal sulcus, STS), are less strongly activated by animated cartoon stimuli than by live action agents performing the same actions.

We conclude then, that given the specific articulations as to the way that automatic sub-processes have been proposed to function in models of belief-desire reasoning (e.g. that such processing occurs in response to the presence of certain stimulus conditions), it may be premature to have diagnosed the non-automaticity of 'belief reasoning' on the basis of the evidence reported in Apperly et al. (2006), and indeed in other recent studies (Liu et al., 2004; Saxe et al., 2006). However, we also consider it premature to consider the evidence presented in the current experiments as evidence 'for' the automaticity of belief reasoning.

This stems first from recognition of the likelihood that 'belief reasoning' is not usefully investigated as a single homogenous process, and might admit of elements, some, but not all of which, might have the signature of automaticity (e.g. Leslie et al., 2004). Second, we suggest that progress might be made via reframing the question from one about whether a cognitive process 'is' or 'is not' automatic and obligatory, to one in which we seek instead to delineate *what are the stimulus conditions that might lead to mental state processing with the signature of automaticity* (as characterized here) and what stimulus conditions result in the requirement that participants take on more controlled or 'top-down' processing strategies.

[6] The test questions included text (for the reality judgment, "Really, where is this?" and for the think judgment, "Where does Garfield think this is?") followed by a picture of the stimulus defining the content about which the reality or belief question is asked (one of two animals that on 75% of trials moves to another box).

Ackowledgements

We would like to thank Ian Apperly and two anonymous reviewers for comments on a previous draft of this manuscript.

References

Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. *Cognition, 106*, 1093–1108.

Apperly, I. A., Riggs, K. J., Simpson, A., Samson, D., & Chiavarino, C. (2006). Is belief reasoning automatic? *Psychological Science, 17*, 841–844.

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence on social perception and cognition. In J. Uleman & J. Bargh (Eds.), *Unintended thought*. New York: Guilford.

Calder, A. J., Lawrence, A. D., Keane, J., Scott, S. K., Owen, A. M., Christoffels, I., et al. (2002). Reading the mind from eye gaze. *Neuropsychologia, 40*, 1129–1138.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*, 531–534.

German, T., & Hehman, J. A. (2006). Representational and executive selection resources in 'theory of mind': Evidence from compromised belief-desire reasoning in old age. *Cognition, 101*, 129–152.

German, T., Niehaus, J. L., Roarty, M. P., Giesbrecht, B., & Miller, M. B. (2004). Neural correlates of detecting pretense: Automatic engagement of the intentional stance under covert conditions. *Journal of Cognitive Neuroscience, 16*, 1805–1817.

Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society, 358*, 549–559.

Kingstone, A., Smilek, D., Ristic, J., Friesen, C. K., & Eastwood, J. (2003). Attention Researchers: It's time to look at the real world! *Current Directions in Psychological Science, 12*, 176–184.

Leslie, A. M., Friedman, O., & German, T. (2004). Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences, 8*, 528–533.

Liu, D., Sabbagh, M. A., Gehring, W. J., & Wellman, H. M. (2004). Decoupling beliefs from reality in the brain: An ERP study of theory of mind. *Neuroreport, 15*, 991–995.

Masconi, M. W., Mack, P. B., McCarthy, G., & Pelphrey, K. A. (2005). Taking an ''intentional stance" on eye-gaze shifts: A functional neuroimage study of social perception in children. *NeuroImage, 27*, 247–252.

Mar, R. A., Kelley, W. M., Heatherton, T. F., & Macrae, C. N. (2007). Detecting agency from the biological motion of veridical versus animated agents. *Social Cognitive and Affective Neuroscience, 2*, 199–205.

Model Penal code § 2.01 (1962).

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*, 255–258.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition, 7*, 17–42.

Roth, D., & Leslie, A. M. (1991). The recognition of attitude conveyed by utterance: A study of preschool and autistic children. *British Journal of Developmental Psychology, 9*, 315–330.

Saxe, R., Schulz, L. E., & Jiang, Y. V. (2006). Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. *Social Neuroscience, 1*, 284–298.

Uleman, J. S. (1989). Framework for thinking about unintended thoughts. In J. Uleman & J. Bargh (Eds.), *Unintended thought*. New York: Guilford.

Uleman, J. S., Saribay, S. A., & Gonzales, C. M. (2008). *Annual Review of Psychology, 59*, 329–360.

Vandenbeld, L. A., & Rensink, R. A. (2003). The decay characteristics of size, color, and shape information in visual short-term memory. *Journal of Vision, 3*, 682.

Wertz, A. E., & German, T. (2007). Belief-desire reasoning in the explanation of behavior: Do actions speak louder than words? *Cognition, 105*, 184–194.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*, 1–34.