The
British
Psychological
Society

www.wileyonlinelibrary.com

A cue-based approach to 'theory of mind': Re-examining the notion of automaticity

Tamsin C. German* and Adam S. Cohen

Department of Psychological and Brain Sciences,
University of California, Santa Barbara, California, USA

The potential utility of a distinction between 'automatic (or spontaneous) and implicit' versus 'controlled and explicit' processes in theory of mind (ToM) is undercut by the fact that the terms can be employed to describe different but related distinctions within cognitive systems serving that function. These include distinctions in the underlying cognitive systems, processes, or representations involved in ToM, distinctions among methodologies or task procedures used to measure ToM, and distinctions among behavioural signatures evaluated as evidence for the engagement of ToM. We propose an approach in which rather than continued dispute over whether or not ToM 'is' or 'is not' automatic, researchers focus instead on discovering what the range of stimulus conditions and task contexts are that give rise to various signatures of the ToM system. These input–output relations will constrain theorizing about the kinds of representations employed, the types of processing operating over those representations, and the overall architecture of ToM mechanisms.

Is human vision automatic? When it comes to 'seeing the world', it certainly seems intuitively to be the case that there is not much overt strategic control required, or even choice in the matter; given certain stimuli and provided we open our eyes, we see. But on a more nuanced analysis, there is good evidence that vision, strictly speaking, is impossible without processes that can be characterized as akin to inferences, albeit unconscious inferences (Gregory, 1997; see also Scholl, 2005). We also know that vision is hugely complex and comprises multiple functionally and even anatomically distinct sub-processes. Most relevant here, however, is the role of *input* or *stimulus conditions*, which are the minimum conditions or thresholds that a visual stimulus must exceed to trigger the visual system. To formalize the notion of automaticity, we must delineate the exact structure of the set of stimulus conditions that apply, and it is critical that we recognize the possible complexity of such input conditions, even when we restrict analysis to just one sub-process.

A good example comes from scotopic vision (the system of rods largely responsible for low-light vision). Even an issue as simple as whether or not light is detected involves

*Correspondence should be addressed to Tamsin C. German, Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106–9660, USA (e-mail: german@psych.ucsb.edu).
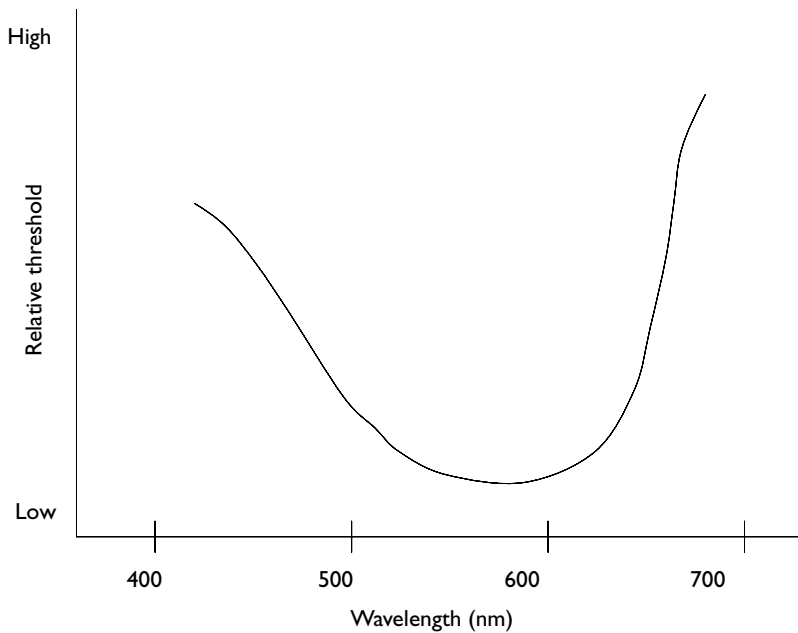
Figure 1. A scotopic vision function showing how the threshold for light detection varies according to the wavelength of light.

the interaction of a number of factors, including (but not limited to) the wavelength of the light and its intensity (Wald, 1945). At intermediate wavelengths of light in the visible range, less light is needed for detection than at the longest or shortest wavelengths, reflecting the fact that different intensities of light, depending on wavelength, are necessary for detection in the same individual (see Figure 1).

In short, the front end of the visual system, just one part of a highly fractionated and complex computational system, has a set of input conditions necessary to engage its function. If you do not meet the input conditions, you do not get in. Abundant research in vision has painstakingly characterized the input conditions of and interactions between dozens of the sub-processes that comprise the various parts of the visual system (see, e.g., reviews in Kingdom, 2011; Morgan, 2011; Shapley & Hawken, 2011; Ungerleider & Bell, 2011; Wandell & Winawer, 2011). It becomes immediately clear that deciding whether the sub-process handling scotopic vision 'is' or 'is not' automatic relies on too crude a distinction; it misses consideration of the myriad of cues that confront the system, some combinations of which are sufficient for its activation and some of which are not.

Now consider a higher order capacity such as 'theory of mind' (hereafter, 'ToM'). We know that some ToM reasoning occurs in response to verbal prompting. But are there other conditions under which information about mental states is encoded, and inferences drawn? Perhaps mechanisms within the ToM system, such as those within vision, can be engaged in response to certain combinations of stimuli that match or exceed input conditions. The case of vision should remind us first that ToM might comprise a number of interacting sub-systems, and second that for each we are likely to uncover a complex interaction between factors when attempting to characterize their input conditions.

In the ToM literature, theorists (e.g., Back & Apperly, 2010) have attempted to determine if cognitive systems might be automatic, spontaneous, or controlled. Rather

than categorize ToM inferences as a whole into one or other category, our approach will be to present three case studies in which a particular signature of ToM functioning is considered *in terms of cues that engage it*. For our purposes, cues include not just the stimuli themselves, but also task contextual information (e.g., repeated exposure to certain unpredicted probes about mental states [see Apperly, Riggs, Simpson, Samson, & Chiavarino, 2006; Cohen & German, 2009], which might induce participants to calculate mental states even under no overt instructions to do so), as well as overt, verbal instructions provided to participants to reason about mental states. It is of course possible that a given task might engage multiple ToM processes and a given ToM process be engaged by a number of cues drawn from across this range of possibilities. Attempting simply to categorize ToM as automatic, spontaneous, or controlled in its function will likely fail to provide most illumination on this complex question.

The indexes of the functioning of ToM we consider are (1) responses to belief probes under 'covert' or 'incidental' conditions (e.g., Apperly *et al*., 2006; Back & Apperly, 2010; Cohen & German, 2009, 2010), (2) participants' looking behaviour while solving overt or covert belief inference problems (e.g., Cohen, Liu, Bernstein & German, 2011; Senju, Southgate, White, & Frith, 2009), and (3) neural signatures (e.g., functional magnetic resonance imaging [fMRI], positron emission tomography [PET], and electroencephalographyn [EEG]) indicating engagement of the 'ToM neural system' (Frith & Frith, 2006).

Case study 1: Responses to unpredictable probes about belief

Apperly *et al*. (2006) presented a test of automaticity in belief reasoning developing the following logic: if people automatically track people's beliefs, then information about those beliefs should be readily available when probed for unpredictably, even if it is irrelevant to the participant's current task.

To test this, Apperly *et al*. created an 'incidental' false belief task where participants performed an explicit task to keep track of an object across a series of events (see Figure 2, top row). The events comprised a false belief scenario, where a woman looked into two opaque containers before placing a marker on one, to signal the location of the object participants were to track. The woman departed, whereupon her companion switched the location of the object, either by moving the object between the containers or by swapping the positions of the containers. The woman returned, at which point an unpredictable probe question appeared asking about the location of the object ('reality' probes), the woman's belief about the object's location ('belief' probes), or, most frequently, something else unrelated to the actual object location or the woman's belief about it ('filler' probes). At the end of the trial, the participants solved their overt task by indicating the location of the object.

Participants' response times to the reality probes were used as a benchmark. If beliefs are tracked automatically, then there should be no difference between response times to the belief and reality probes. A delay in responding to the belief probes (relative to the reality probes) was taken to signal that belief information had not been encoded in the cognitive system in response to the woman's communicative cue (placement of the marker indicating the object's location) and had to be calculated only when the probe appeared.

A key feature of this procedure is that the probe questions about belief and reality are not predictable in advance, lest participants begin to track belief due to contextual factors, recognizing as the experiment progresses that belief is repeatedly enquired
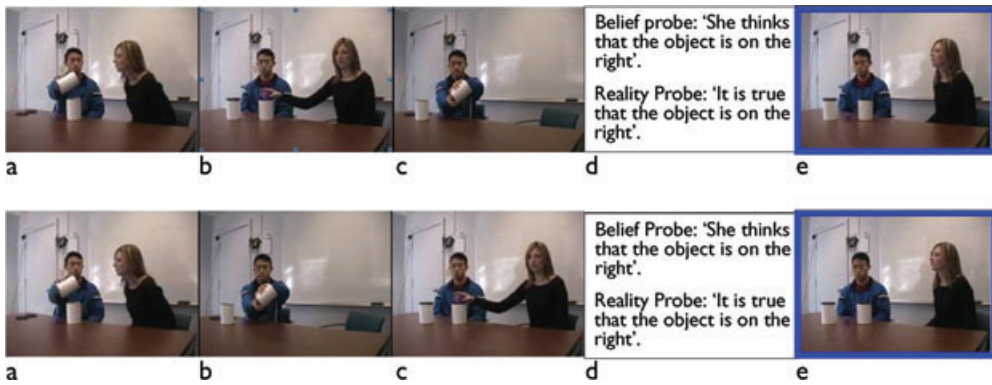
Figure 2. The sequence of events in Apperly *et al.*, 2006 (top row). (a) The woman looks into the containers, (b) gives the signal, (c) the man swaps the containers (invisible trials) or the location of the object (visible trials), (d) the probe appears, and finally (e) the blue frame appears indicating the end of the video. The bottom row shows the modified 'short delay' condition (Cohen & German, 2009). (a) The woman looks into the containers, (b) the man swaps the containers (invisible trials) or the location of the object (visible trials), (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).

after.[1] To prevent this, Apperly *et al*. embedded the test questions among a larger number of filler probe questions seeking information about other aspects of the scenarios.

Apperly *et al*. observed that responses to belief probes were indeed slower than those to reality probes in the key 'incidental' belief condition, a result consistent with belief information not having been encoded in response to stimulus information (the woman marking the object's location). Moreover, under conditions where participants were asked to track belief information (either alone or along with reality information), the difference between the belief and reality probes vanished, suggesting that instruction to track belief is sufficient for belief information to be readily available in the cognitive system.

One weakness of the Apperly *et al*. method is that belief probes occur considerably after the presentation of information relevant to its encoding (the woman's communicative signal). The woman's action occurs early in the sequence of events, and her belief is rendered false only when the object changes locations later. The probes come after the change of object location, and to be available at the probe, information about belief must be encoded and maintained in the cognitive system for at least 20 s. This leaves open the possibility that information about belief is encoded in response to the cues, but decays before the probe appears.

Cohen and German (2009) removed the requirement for belief information to be maintained over time by switching the sequence of events such that the object's locations were switched after the woman had looked inside the containers, but before she placed the marker (see Figure 2, bottom row). This way, probes about belief follow the relevant communicative cues with a much shorter latency. Under these 'short delay'

---

[1] *Contextual triggering (e.g., due to frequency of an event or probe) of belief tracking is neutral with respect to whether or not the participant is aware that he/she is processing belief.*

conditions, reaction times to belief probes were faster than those to the reality probes, indicating belief information is encoded in response to the cues provided in the woman's communicative signal. Cohen and German argued further that in the Apperly *et al*. 'long delay' condition, information that may have been encoded in response to the cues decayed before the probes arrived, given no reason for its maintenance.

In a follow-up to Apperly *et al*. (2006), Back and Apperly (2010) provide a further test for non-automaticity from conditions using long delays in their paradigm, adding a true belief condition alongside the false belief condition. They replicated the slower responses for belief probes seen in Apperly *et al*. (2006; Cohen & German, 2009, long delay condition), but also showed that performance for reality probes in false belief conditions was slower than that for true belief conditions. They attributed this to interference between belief and reality information in the false belief condition. But how did belief information get into the system if not in response to the woman's communicative cue? Back and Apperly argue that the belief information was 'spontaneously' rather than automatically encoded, a distinction drawn from work on text processing: '[these inferences] are not viewed as automatic because their processing is contingent on a variety of contextual factors (Back & Apperly, 2010, p. 55)'.

By 'contextual factors', Back and Apperly (2010) seem to suggest the possible cuing of thinking about beliefs in response to repeated belief probes across the course of the experiment. While belief inferences in Cohen and German (2009) and the interference effects in Back and Apperly might have resulted from such contextual effects, a test of this idea in Back and Apperly (2010, Exp. 3) renders results that are equivocal, in our view.

A more serious concern is that, given the foregoing, it is unclear that *any* pattern of results would comprise evidence for 'automatic' belief inferences in this paradigm. The actions of the character are identical in the short versus long delay conditions, as is the ratio of test trials to filler trials. If participants are induced into generating 'spontaneous' inferences by the woman's communicative act under short delay conditions,[2] or form a task-set to track belief given the context of repeated belief probes, then presumably these factors could also account for fast belief inferences in long delay conditions (had they been observed in the first place).

It is not really clear, therefore, that the unpredictable probe methodology can shed any light on the question of automaticity, at least when considered as a binary yes/no question. However, under the cue-based approach advocated here, this set of studies has provided important information about the conditions under which ToM mechanisms are engaged, and the fate of the information that is encoded:

(1) Seeing overt communicative gestures can lead to the generation of information about a social agent's belief, even when considering mental states is irrelevant to the participant's overt task.

(2) Such belief encoding is seen via two different indexes; response times to incidental belief probes that are *at least as fast* as those to (overtly tracked) reality probes (Cohen & German, 2009) and/or interference with explicitly tracked reality information when beliefs are false (Back & Apperly, 2010).

---

[2] *Back and Apperly (2010) suggest that the woman's (incorrect) marking of the container, immediately followed by the presentation of the probe, might have invited participants to make spontaneous inferences in order to form a coherent interpretation of the woman's behaviour.*

(3) Beliefs might be encoded as a result of context that builds up over repeated belief probes (either with or without awareness or top-down control, see footnote 1).

This leads to a testable prediction. Specifically, if encoding is the result of some implicit task set generated by a context of repeated belief probes, then it should emerge only later in the experiment, and certainly not be present on the first trial. We tested this hypothesis by revisiting the Cohen and German (2009) data and performing a Bayesian analysis (Gallistel, 2009) of the first trial data only. This analysis resulted in odds favouring the null hypothesis over the alternatives of about 6:1. That is, the most likely hypothesis is that belief inferences and reality inferences do not differ. This suggests that the belief encoding is at least as efficient as the tracked reality information from the outset, a result incompatible with the idea that a task set builds up (tacitly or explicitly) over repeated exposure to belief probes.

Case study 2: Spontaneous looking as an index of belief inference

Clements and Perner (1994) showed that 3-year-old children unable to answer a standard false belief action prediction question nonetheless looked in anticipation towards the location where the character would return. The publication of that study provoked the development of the idea of a distinction between 'implicit' and 'explicit' knowledge in ToM, where anticipatory looking was considered a possible window into earlier, or even representationally different knowledge of ToM.[3]

We focus here on recent investigations of eye tracking and anticipatory looking in adult populations, where we are concerned with applying a cue-based approach, and finish this section by touching on how to apply the approach to the infant literatures on 'violation of expectation' (Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007) and anticipatory looking (Southgate, Senju, & Csibra, 2007). We hope that the approach advocated here may help to illuminate those results as they continue to emerge.

Eye gaze is a sensitive behavioural measure and certain patterns of looking can be interpreted as signaling the engagement of the ToM system. But a careful analysis of the cue conditions that determine whether a given looking pattern occurs is still required. Firstly, note that the initial demonstrations of anticipatory looking occurred in response to a verbal prompt: 'I wonder where Sam will look for his cheese', which albeit not a direct question, might be shown to be a necessary cue for the engagement of the behaviour.

More recent investigations of younger children show their anticipatory looking reflects belief tracking (Southgate *et al*., 2007), while adults with an Autism Spectrum diagnosis who show good performance on overt verbal measures of false belief do not show such anticipatory looking (Senju *et al*., 2009). But what is the status of these phenomena with respect to the question of the cues that engage the ToM system? Such results are often characterized as demonstrations (or an absence) of 'spontaneous ToM' (see e.g., Senju *et al*., 2009), since they do not use a verbal prompt to invoke looking. However, they in fact do employ familiarization/training in which the participants learn to expect a light/chime to occur just before the social agent returns, and this set of cues could prove critical.

---

[3] *We defer to the conclusion a comment on the other ways in which the implicit–explicit distinction can and has been deployed within the ToM literature.*

Besides the possible cuing/elicitation of ToM inferences via a verbal (i.e., 'I wonder where Sam will look for his cheese') or other prompt (i.e., a light/chime indicating the imminent return of a protagonist), the possibility remains that the generation of belief inferences that drive looking results solely from the participant having being exposed to an unfolding event involving social agents and beliefs. Stronger evidence for the role of the cues in the event itself would come from situations where the looking pattern emerges while participants are engaged in a different task (as described in the covert belief studies in the previous section), and where belief information is of no help to the participant for that task, or where such information might even interfere with their overt task (see e.g., Samson, Apperly, Braithwaite, & Andrews, 2010).

We recently investigated this possibility (Cohen *et al.*, 2011). Participants were presented with a series of animations depicting a typical object transfer scenario, in which the character either leaves the room (false belief) or stays inside (true belief). Having watched the animation, a fixation screen appears for a few seconds, after which the character is shown standing either behind the location containing the object or behind the empty location. The participants' only task is to indicate whether the character is on the left or right. The character's location is random, and so the preceding events have no bearing on performing the task. We analysed a 4-s window prior to the final outcome screen for signs of looking guided by any of the features represented in the videos.

Participants' looking during this window was driven in part by the location of the object; there was a bias to spend more time looking where the object was than at the empty location. However, the character's belief also influenced the participants' looking. In the true belief condition, nearly all looking across the two locations was directed at the location containing the object, but there was a significant reduction in this bias in the false belief condition.

Note that tracking either the object or the belief is of no use to the participants in solving the task, and so this finding goes further than existing demonstrations in showing that properties of the stimulus events can cause engagement of ToM systems as measured by this looking pattern index. Further discussion of the relevant stimulus events, identification of which is crucial to fleshing out the cue-based approach, is taken up in case study 3.

How does the encoding approach apply to the developmental literature? One puzzle that the approach may shed light on is the discrepancy between belief-based anticipatory looking in 2.0-year-olds (Southgate *et al.*, 2007) but not in older 2-year-olds (Clements & Perner, 1994). One hypothesis for this discrepancy is that in the task with older 2-year-olds, the target object was present, placing increased demands on inhibition, whereas in the task with the 2.0-year-olds, the target object was absent (Southgate *et al.*, 2007). Combining that hypothesis with a cue-encoding approach, it is possible that when the target object is present, it engages other systems (i.e., object tracking for fixing and updating belief) that compete with belief tracking for control of eye movements. Another possibility is that the presence of the target acts as a cue that directly engages mentalizing, and because of default settings in the ToM system, results in a true belief attribution (see e.g., Leslie, Friedman, & German, 2004). Successful performance comes from inhibiting this prepotent, default response (Leslie & Polizzi, 1998).

In a critical test of the inhibition hypothesis, Wang, Low, Jing, and Qinghua (in press) show that when 3-year-olds are divided into younger and older groups, there is an effect of object salience: younger 3-year-olds are less likely to show correct anticipatory looking (hereafter, AL) when the target is present than when absent. They also show that the

mean looking time to the correct location is significantly greater in the target-absent than target-present condition, for both 3- and 4-year-olds (although the reported effect size is small, it folds the 3- and 4-year-olds together; casual inspection suggests the effect size should be larger for the 3-year-olds – nearly double that of the 4-year-olds). Although the authors interpret their data as not supporting the Southgate *et al*. inhibition hypothesis due to other analyses suggesting a weak or null effect of object salience, the results appear mixed, with at least some support for the role of inhibition.

We think an even stronger test of Southgate *et al*.'s claim would be to run the target present–absent manipulation in 2-year-olds since the hypothesis makes the strongest predictions at that age: the 2-year-olds should show correct AL when the target is absent, but not when it is present. Since 3-year-olds already show correct AL when the target is present, one certainly expects, on an inhibition account, that they will do at least as well when the target is absent. But there is not much room for improvement because performance is already near ceiling when the target is present. Not finding a difference when target is absent or present in 3-year-olds is plausibly due to ceiling effects (in fact, the inhibition account predicts that at some point, when inhibitory control is sufficiently developed, AL should not differ as a function of object salience). It is striking that differences still emerge, consistent with Southgate *et al*.'s hypothesis, when the 3-year-olds are subdivided into younger and older groups. Because of these issues, 2-year-olds are the critical test group, and for now, the inhibition hypothesis is still on the table, if not bolstered by the results from the 3-year-olds.

Another consideration is that when children are presented with the same stimulus cues, but in one case generate incorrect verbal response and in another case generate correct non-verbal responses (in violation of expectation [hereafter, VOE] or AL tasks), divergent performance cannot be explained by appealing to the cues since they are the same. Instead, the difference appears to reflect differences in the access that the output systems have to the representational products of the belief inference pathway. The fact that 2- and 3-year-olds generate correct looking responses in non-verbal tasks suggests that sufficient cues are available to engage belief processing. Since they fail to give correct verbal responses but make correct eye movement responses in VOE and AL non-verbal tasks, this suggests that non-verbal output systems might have privileged access to the mechanisms generating belief representations (relative to the verbal output system). In fact, children's performance is reminiscent of a functional disconnection syndrome in which belief information appears to be in the ToM system, but either (1) fails to be routed to language production devices or, (2) is corrupted during transmission, or finally (3) does not arrive in the language output system in time to affect the verbal response (i.e., information about belief might be slower than information about where the object actually is to arrive in language output systems).

In cases where there is successful performance in VOE or AL tasks, it suggests that researchers have stumbled across sufficiently relevant inputs that can engage the ToM system, meeting its input conditions. VOE tasks are interesting because the final search frame itself might be a potent cue to an agent's belief, which is consistent with previous work suggesting that action towards objects leads people to generate mentalistic explanations (Wertz & German, 2007). Using a VOE paradigm and a change detection task, Kovács, Téglás, and Endress (2010) introduced an agent who observes a hiding event and show that infants look longer when their own beliefs or the agent's beliefs about the location of the object are violated. These effects did not hold if the agent was substituted with a pile of boxes (Kovacs *et al*., 2010). This suggests that the presence of an agent is one cue critical to meeting the input conditions for triggering

ToM processing. This style of research is precisely what we have in mind when one takes a cue-encoding approach. The key step for both infant and adult researchers is to characterize the matrix of inputs that engage the ToM network and how it might change over development.

Case study 3: Activation of the ToM neural signature

Studies seeking to identify brain areas involved in ToM tasks, mostly via PET and fMRI, as well as a handful using event-related potentials (ERPs; Liu, Sabbagh, Gehring, & Wellman, 2004; Sabbagh & Taylor, 2000) have utilized a wide variety of comparisons between mental state and control content, including 'ToM' vignettes versus 'physical causal' stories or cartoons (Fletcher *et al*., 1995), describing random, versus mechanical versus mentalistic animation sequences (Castelli, Happé, Frith, & Frith, 2000), playing a strategic game against an imagined person versus computer (Gallagher, Jack, Roepstorff, & Frith, 2002), viewing of 'pretend' versus real actions while performing an unrelated task (German, Niehaus, Roarty, Giesbrecht, & Miller, 2004), and analysis of activations caused by different kinds of eye gaze (Calder *et al*., 2002).

Across this wide range of comparisons, a suite of areas is shown to be more active in ToM tasks than in control tasks. These areas include the medial prefrontal cortex, the temporal poles and the temporal parietal junction (TPJ; see Frith & Frith, 2006).The extent of work using these methods allows us to utilize activation of the ToM network in functional imaging as a signature of the engagement of ToM processes, and apply a cue-based approach to determine under what conditions this network can be activated.

Early work on the neural signal of ToM relied exclusively on tasks with overt instructions to participants to think about mental state content or to make strategic responses about the actions of social agents. For example, Fletcher *et al*. (1995) explicitly instructed subjects to reason about the mental states of the protagonists in ToM stories, and explicitly directed them to *avoid* considering mental states for control tasks. Castelli *et al*. (2000) informed subjects that they would see three different types of animation, and that some would seem 'as if [the triangles] were taking into account their reciprocal feelings and thoughts' (Castelli *et al*., 2000, p. 322).

German *et al*. (2004) considered the neural signal associated with ToM where there were no explicit task instructions, looking at brain activation patterns in response to pretend actions versus real actions. Participants were engaged in a different overt task (making judgments about whether the video clips finished prematurely or not). The results showed the ToM neural signature in response to the cues of pretending (Richert & Lillard, 2004), even under covert conditions. Unusual or atypical actions *are not required*, however; passive viewing of video sequences involving social interactions also engages the ToM neural network in the absence of task instructions (Mar, Kelly, Heatherton, & McCrae, 2007). As Mar *et al*. (2007 , p. 204) put it: 'the activations found by previous researchers are not simply the result of instructions directing participants to engage in mentalizing tasks, but that these regions of the brain are also employed when individuals infer mental states in a natural, non-directed and more ecologically valid manner'.

But what exactly are the cue structures that result in engagement of the ToM system? One obvious possibility is that where participants passively view streams of behaviour (e.g., Mar *et al*., 2007), they overtly and consciously think about the thoughts and other mental states of the actors despite no instructions to do so. This seems plausible in cases where there is no other task set, and it may be the case also in studies where participants

are instructed to perform a different overt task (German *et al.*, 2004). Alternatively, perhaps the engagement of ToM occurs in response to the behavioural cues unfolding in the stimulus stream without conscious internal consideration of mental states. These studies provide no evidence either way.

The role of stimulus cues versus overt instruction (either experimenter or participant generated) has been illuminated in other work. For example, Saxe, Schulz, and Jiang (2006) argue that strategic mentalizing is required to see engagement of ToM areas. They presented simple animations to participants and required them to make responses under two different tasks instructions. In one condition, participants responded following a simple arbitrary rule based on the spatial orientation of one of the figures in the animation. In the other, they were instructed to use mental state reasoning to respond.

Saxe *et al.* observed a blood oxygen level dependent (BOLD) response in areas most reliably activated by ToM inferences (right temporo-parietal junction [rTPJ] and, less strongly, left TPJ) only when participants construed the scenario mentalistically. When using the non-mentalistic algorithm, there was a suppressed response. The authors interpreted this result as evidence against 'automatic' engagement of ToM since the stimuli presented in each instruction condition were identical but the instructions differed, which appeared to influence how participants construed the stimulus (Saxe *et al.*, 2006, p. 295).

This result raises a number of interesting issues on a cue-based approach. First, setting aside the question of task instructions, to what extent is the *richness* of the stimuli involved in a given task an important consideration? The animations used by Saxe *et al.* comprised simple line drawings of stick figures, and one possibility is that such impoverished cues are less likely to overlap the input conditions for engagement of the ToM system.

Mar *et al.* (2007) provided evidence suggesting that the richness of the stimuli matters. They showed participants clips in which the same set of live actions that invoked the ToM neural signal (see above) were rendered as animations. The animations were created directly from the live video sequences, and therefore retained identical motion patterns and interactions between agents. Under such minimally degraded cues, the response in TPJ was considerably attenuated, showing only about 50% of the signal increase seen in response to regular videos. This suggests that the ToM system's activation is greatly influenced by the richness of the behavioural cues available to it.

The context of task instructions is another relevant cue to consider. Animated sequences involving sets of simple geometric shapes (e.g., 'self-propelled' triangles) have been shown to have the capacity to engage mental state reasoning under conditions of overt instruction (Castelli *et al.*, 2000). While the stimuli can be considered impoverished in terms of their morphological features, they engaged in rich patterns of contingent interaction (coaxing and teasing), and even very simple contingent interactions can be sufficient to drive early manifestations of social behaviour in infants (e.g., gaze following, Johnson, 2003). Beyond a general instruction to explain the events, however, more detailed cuing of participants prior to each animation that the sequences 'would involve feelings and thoughts' did not change the extent of activation of ToM areas.

Returning to Saxe *et al.*, (2006), we cannot know whether the animations used were capable of engaging ToM areas absent any instruction to overtly mentalize, because there was no 'passive viewing' condition. Even if the animations afforded mental state interpretation without instruction (unlikely in the light of Mar *et al.*, [2007]), it is still possible that *imposing overt instructions to use a non-mental spatial algorithm* might have shifted participants' attention away from belief relevant features of the stimulus,

thereby suppressing the ToM neural signature. The signal change in rTPJ under arbitrary rule instructions did indeed consist in a *reduction* from baseline (Saxe *et al*., 2006, Figure 4, p. 293).

One limitation stemming from the use of fMRI and PET activation patterns as indexes of ToM function, under a cue-based approach, is that such signatures rely on a sluggish response of blood flow in the brain. As a result, the temporal relationship between presentation of cues relevant to the engagement of ToM processing and the BOLD response that arises is hard to determine precisely. Studies separating the pattern of activation occurring during stimulus presentation (e.g., written stories describing people engaging in actions) from activations occurring in response to participants' explicit task performance (e.g., Aichhorn *et al*., 2009) can shed light on this question. Aichhorn *et al*. show that activations in rTPJ happen prior to the point at which any action predictions are required, an outcome that is consistent with generation of belief inferences in response to either the stimulus information or as a result of overall task context (in which repeated questions about belief may be expected).

ERPs, measured via EEG, provide evidence with far greater temporal resolution than fMRI and PET (at the price of greatly reduced spatial resolution) and are especially well suited for the consideration of stimulus-driven processing.

Liu *et al.*, (2004) had participants watch a series of cartoon animations and then make judgments about where a character thought an object was located or where the object was actually located. They reported an ERP component for ToM questions separating from the reality question component questions about 800 ms post-stimulus, arguing that this peak was too late to signal automatic, 'perception like' processes (Liu *et al*., 2004, p. 995).

The ERP in this particular study, however, was actually locked to presentation of a picture of a stimulus defining the content about which either a belief or reality question had already been asked. This stimulus was presented subsequent to the question itself, and subsequent too, to events over which belief and reality calculations were made. Therefore, the study is really measuring the on-line neural signal associated with an overt and instructed belief judgment rather than the possible on-line belief processing associated with the unfolding event itself.

What is needed is evidence of on-line processing of beliefs in response to stimuli where there are no overt questions about belief, and no context in which a task set emerges inducing participants to adopt a strategy to calculate belief. Ideally, as noted before, such a strategy should not help participants with the overt task they are performing.

Liu (2011) provided such evidence using the same task and stimuli described in case study 2 above where participants watched animations of characters engaging in true and false belief event sequences. The participants' overt task was simply to indicate, after the animation, when the character was standing behind the right location. This was therefore purely a spatial task; participants could solve it by ignoring the preceding animation entirely.

Liu's study leveraged the P300 (an ERP component associated with infrequent, unexpected, or inconsistent events) and arranged the trials such that the target standing on the right was a low-frequency occurrence (20%). Of the remaining trials (80%), where the character appears on the left, some (another 20% of the total trials) occurred when the events in the story had unfolded such that the character searched in an unexpected location, and the remainder occurred when the character searched in the expected location. The logic of the study was that in addition to the P300 observed to the

low-frequency targets (i.e., character standing on the right), there might be a similar result when the character was seen in a position that violated his or her belief (unexpected trials). Note that this could only occur if belief had been inferred in response to the stimuli during the animation.

Liu's results showed two early components of the P300 (P3a and P3b) were greater for the belief violating trials (in both true and false belief cases) than for 'expected' trials, suggesting that the ToM system is tracking beliefs, even where it is of no value for performing the task. Strikingly, this pattern was maintained over hundreds of trials, consistent with the mandatory calculation of belief in response to cues presented to the system.

Conclusions

Taking a cue-based approach to the question of the engagement of different signatures of the ToM system provides rich data that can inform theories about the structure, representations, and processes within ToM. In our view, the utility of the distinction between, variously, 'implicit' and 'explicit', 'automatic' (or 'spontaneous') and 'controlled', 'exogenous' versus 'endogenous', 'bottom-up' versus 'top-down', and so on, is undercut in work on ToM, just as it has been in other areas of psychological research (such as memory), because it is not always clear where and how the distinction should be applied.

A distinction might apply at the level of 'task' (e.g., overt/elicited belief questions vs. unpredictable/covert belief probes while performing a different task), at the level of 'behavioural measure' (e.g., verbal responses to a question about belief vs. spontaneous looking in response to an 'I wonder where' prompt), or finally at the level of a theoretical claim about different kinds of process, representation type or mechanism within the cognitive system (as in 'does eye gaze indicate implicit knowledge of theory of mind' [Clements & Perner, 1994] or 'do humans have two systems to track beliefs and belief like states' [Apperly & Butterfill, 2009]).

We have suggested in this paper that researchers might usefully begin by drawing mappings between precisely defined stimulus 'syndromes' and observable signatures of ToM function. It is important to stress, however, that we are not advocating a retreat towards 'behaviourism'. We proceed from the view that by studying the cue structures that engage a system, we are discovering something about *input conditions* of that system. Input conditions are necessarily 'inside the mind'.

We fully endorse theorizing about structure within the ToM cognitive system. We have ourselves proposed theoretical ideas on the internal structure of ToM and the various representational and executive selection resources it might employ (Cohen & German, 2009; German & Hehman, 2006; Leslie *et al*., 2004; Leslie, German, & Pollizi, 2005; Wertz & German, 2007; Yazdi, German, Defeyter & Siegal, 2006). It is likely that the ToM system will be shown to comprise a number of sub-systems and involve different kinds of representations. Some parts of the system may one day be usefully characterized as implementing 'implicit knowledge' and others 'explicit knowledge'. For now, however, we proceed in recognition of the fact that available evidence vastly underdetermines the possible theories that might be true. Focusing on obtaining more information relevant to the task of mapping observable stimulus complexes to known signatures of ToM function will be of great value, constraining and narrowing the space for specific theoretical proposals.

Acknowledgements

References

Aichhorn, M.,, Perner, J., Weiss, B., Kronbichler, M., Staffen, W., & Ladurner, G. (2009). Temporo-parietal junction activity in theory-of-mind tasks: Falseness, beliefs, or attention? *Journal of Cognitive Neuroscience*, *21*, 1179–1192.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*, 953–970.

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

Back, E., & Apperly, I. A. (2010). Two sources of evidence on the non-automaticity of true and false belief ascription. *Cognition*, *115*, 54–70.

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., & Young, A. W. (2002). Reading the mind from eye gaze. *Neuropsychologia*, *40*, 1129–1138.

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, *12*, 314–325.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*, 377–395.

Cohen, A. S., & German, T. C. (2009). Encoding of others' beliefs without overt instruction. *Cognition*, *110*, 356–363.

Cohen, A. S., & German, T. C. (2010). Reaction time advantages for calculating beliefs over public representations signal domain specificity for 'theory of mind'. *Cognition*, *115*, 417–425.

Cohen, A. S., Liu, D., Bernstein, D., & German, T. C. (2011). *Eye movements and reaction time reveal automatic belief computation under minimal cue conditions*. Paper presented at the biennial meeting of the Society for Research in Child Development, Montreal, Canada, April, 2011. Manuscript in preparation.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of 'theory of mind' in story comprehension. *Cognition*, *57*, 109–128.

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

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

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

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.

Gregory, R. L. (1997). *Eye and brain: The psychology of seeing*. Princeton, NJ: Princeton University Press.

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

Kingdom, F. A. A. (2011). Lightness, brightness and transparency: A quarter century of new ideas, captivating demonstrations and unrelenting controversy. *Vision Research*, *51*, 652–673.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*, 1830–1834.

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

Leslie, A. M., German, T., & Pollizi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, *50*, 45–85.

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

Liu, D. (2011). *An ERP study of automatic mentalizing*. Paper presented at the biennial meeting of the Society for Research in Child Development, Montreal, Canada, April, 2011. Manuscript in preparation.

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.

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

Morgan, M. J. (2011). Features and the primal sketch. *Vision Research*, *51*, 738–753.

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

Richert, R. A., & Lillard, A. S. (2004). Observers' proficiency at identifying pretense acts based on behavioral cues. *Cognitive Development*, *19*, 223–240.

Sabbagh, M. A., & Taylor, M. (2000). Neural correlates of the theory-of-mind reasoning: An event-related potential study. *Psychological Science*, *11*, 46–50.

Samson, D., Apperly, I. A., Braithwaite, J., & Andrews, B. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 1255–1266.

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.

Scholl, B. J. (2005). Innateness and (Bayesian) visual perception: Reconciling nativism and development. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Structure and contents* (pp. 34–52). Oxford: Oxford University Press.

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger Syndrome. *Science*, *325*, 883–885.

Shapley, R., & Hawken, M. J. (2011). Color in the cortex: Single and double opponent cells. *Vision Research*, *51*, 701–717.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by two-year-olds. *Psychological Science*, *18*, 587–592.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs to 13-month-old infants. *Psychological Science*, *18*, 580–586.

Ungerleider, L. G., & Bell, A. H. (2011). Uncovering the visual "alphabet": Advances in our understanding of object perception. *Vision Research*, *51*, 782–799.

Wald, G. (1945). Human vision and the spectrum. *Science*, *101*, 653–658.

Wandell, B. A., & Winawer, J. (2011). Imaging retinotopic maps in the human brain. *Vision Research*, *51*, 718–737.

Wang, B., Low, J., Jing, Z., & Qinghua, Q. (2011). Chinese preschoolers' implicit and explicit false-belief understanding. *British Journal of Developmental Psychology*. Advance online publication. doi:10.1111/j.2044-835X.2011.02052.x

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