The Cognitive Basis of Mindreading

Ian Apperly

Why did Anna Karenin throw herself under a train? A satisfactory answer to this question will surely refer to Anna's mental states—her thoughts and feelings, desires, and intentions. Most readers of Tolstoy's novel would consider such mindreading essential to understanding the story. They might also find that an important part of Tolstoy's craft is the generation of tension between Anna's perspective and emotional state and those of other characters, and their own perspective as a reader. It is deeply revealing about the nature of mindreading that we find it quite natural to think about the mental states of a fictional character, from a different place and time, in an unusual set of personal circumstances, and this exposes important limitations of common claims and assumptions about mindreading.

Neuroscientific approaches have much to teach us about the nature of mindreading but, as in other areas of cognitive neuroscience, they are at their most powerful when combined with clear hypotheses about the cognitive processes involved. I begin by considering the limitations of some prominent theoretical ideas about mindreading. I will go on to describe a cognitive account that, I think, provides a better foundation for a cognitive neuroscience of mindreading. I will highlight examples of what neuroscientific approaches have already told us about the cognitive basis of mindreading, before considering some exciting future prospects.

Mindreading Mantras

Mindreading has been extensively theorized by psychologists, linguists, and philosophers. This offers a rich inheritance to empirical investigators. However, bold conjectures about how mindreading might work have sometimes become received

I. Apperly (✉)
School of Psychology, University of Birmingham, Birmingham, UK
e-mail: i.a.apperly@bham.ac.uk

M. Gilead, K. N. Ochsner (eds.), *The Neural Basis of Mentalizing*,
https://doi.org/10.1007/978-3-030-51890-5_18

wisdom about how it does work or must work, which can cloud our thinking about what mindreading is and how to study it. To persuade you that it's worth engaging seriously with questions about the cognitive basis of mindreading, let me challenge some oft-repeated claims.

Mindreading is not just "decoding" of mental states from behaviour. It is commonly assumed that mental states can be decoded from behaviour, in much the same way as words can be decoded from text (e.g., Heyes, 2018). Of course it is true that being able to interpret a facial expression as evidence of an emotion, or search behaviour as evidence of a belief about an object's location and a desire to find it, are important components of mindreading. Equally, however, such decoding is not the essence of mindreading. It is clear from the example of Anna Karenin that we may mindread without direct perceptual access to behaviour. Moreover, many of the mental states we might ascribe to Anna—such as her anxiety about her social position—follow from facts about her background, about other characters, or the context, none of which we have observed. Mindreading real people is no different. Moreover, Tolstoy sometimes simply tells us what Anna is thinking; just real people sometimes report on their own mental states, and those of others. Therefore, while observed behaviour is surely one important input for mindreading, it is not necessary and, other than in the simplest cases, it is not usually sufficient.

People do not have a "theory" of mind It is commonly claimed that our mindreading abilities consist in theory, involving concepts—"belief", "desire", "intention", etc.—and principles for how they combine (e.g. Davies & Stone, 1995; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). Just as someone who knows the words and grammar of a language is equipped to parse sentences of that language, so someone with a "theory of mind" would be able to use mental states for explanations or predictions of behaviour. However, unlike linguistic grammars, 30 years of research on mindreading has not codified the supposed principles by which mental states interact for realistic scenarios (Stuhlmüller & Goodman, 2014). There is no extant theory that can parse Anna's circumstances into a reliable set of thoughts and feelings. Instead there are good grounds for supposing that the complexity of the interactions among mental states and between mental states and behaviour is uncodifiable (e.g. Davidson, 1990). Note that this should not be taken as support for "simulation" accounts of mindreading, which do not offer easy solutions to this problem (e.g. Apperly, 2008).

Mindreading does not make unique reasoning demands. An influential early account suggested that mindreading poses unique logical problems, which require a unique representational solution (e.g. Leslie, 1987). This strong hypothesis is difficult to sustain since similar logical problems arise when we need to set aside our own current situation to reason about different times, places, or counterfactuals (Barwise & Perry, 1983; Fauconnier, 1985). Moreover, there are empirical associations between mindreading tasks and non-mindreading tasks that are matched in their logical and structural requirements (Perner & Leekam, 2008). A reasonable conclusion from such work is that mindreading poses some exacting representational challenges, but not unique ones (Apperly, 2010).

Mindreading is not simply automatic What people mean when they claim that mindreading is automatic seems to range from the intuition that mindreading is natural and effortless to a firm commitment to mindreading being a quasi-perceptual Fodor-module (e.g. Leslie, 2005; Stone, Baron-Cohen, & Knight, 1998). Either way, direct investigations have provided evidence that mindreading meets important criteria for automaticity in some circumstances (e.g. Kovács, Téglás, & Endress, 2010; Qureshi, Apperly, & Samson, 2010; Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010; van der Wel, Sebanz, & Knoblich, 2014), but shows clear non-automaticity in others (e.g. Apperly, Riggs, Simpson, Chiavarino, & Samson, 2006). While some of these results remain controversial (e.g. Heyes, 2014; Phillips et al., 2015), it has been suggested that they can be reconciled in a "two systems" account, whereby humans have the capacity to make a minimal set of mindreading inferences automatically, and a second ability that is more effortful but flexible enough to cope with the complexity of full-blown mindreading (e.g. Apperly & Butterfill, 2009; Low, Apperly, Butterfill, & Rakoczy, 2016). The latter would be key to mindreading Anna Karenin, where, on analogy with other inferences made during comprehension, mindreading would be spontaneous (i.e. uninstructed) but conditional on having the requisite processing resources and the motivation to use them (Apperly, 2010).

In summary, mindreading involves much more than "decoding" mental states from behaviour, not least because there is nothing like an exhaustive code. Mindreading makes similar representational demands to structurally similar problems that have nothing to do with mindreading. Only some kinds of mindreading judgement show signs of automaticity; others—such as the problem of figuring out why Anna Karenina threw herself under a train—are clearly effortful and contingent on resources and motivation. In functional terms, such mindreading requires complex, flexible processing over our full database of knowledge about the world, and so fits Fodor's criteria for "central" rather than "modular" processes (Fodor, 1983). From this perspective, it should be no surprise to discover that mindreading involves a rich set of processes for representation, reasoning and control, supported by a network of brain regions. However, this also demands some kind of functional model to organize existing findings and guide new research. Below I summarize such a model. A fuller justification in terms of empirical findings can be found in Apperly (2010).

A Cognitive Model of Mindreading

The great majority of mindreading tasks present a situation involving an agent with mental states that differ from participants'. The agent's mental states must be inferred and either reported, or else used to predict their subsequent behaviour. In doing so these tasks combine and confound most of the functional processes that contribute to mindreading. If we only used this approach, it would be like trying to understand the cognitive and neural basis of language by only ever presenting

participants with tasks that combine every level of phonological, syntactic and semantic processing in a full cycle of comprehension and production. To break out of this problem, we need theories of the functional components of mindreading and tasks that allow putative functional components to be distinguished.

I find box and arrow models extremely useful for organizing ideas about the cognitive basis of mindreading. In the following, I focus on mindreading what someone thinks or knows. The level of description is "computational" in Marr's sense (Marr, 1982), so model components say something about what the system is doing, with no commitment to the algorithmic or neural implementation of those functions. However, such a model of these functions is, I think, essential for explaining or predicting the demands made by different mindreading tasks and how these affect the recruitment of neural systems during mindreading.

The horizontal dimension of Fig. 1 distinguishes the need to infer what someone else is thinking from the need to store this information, and from the use of this information to predict or explain what someone is doing or saying. The vertical dimension distinguishes between "system 1" (below) and "system 2" (above) processes (e.g. Evans & Stanovich, 2013). System 2 processes enable highly flexible mindreading, so in principle I could ascribe to you or to Anna Karenin any thought that I could entertain for myself. However, this flexibility comes at the expense of System 2 thinking making higher demands than System 1 on scarce resources for memory and cognitive control (e.g. Low et al., 2016). System 1 trades reduced flexibility for increased efficiency. Increased efficiency is evidenced in the apparent automaticity with which some mindreading processes occur, and limited dependence on cognitive control processes (e.g. Kovács et al., 2010; Qureshi et al., 2010;

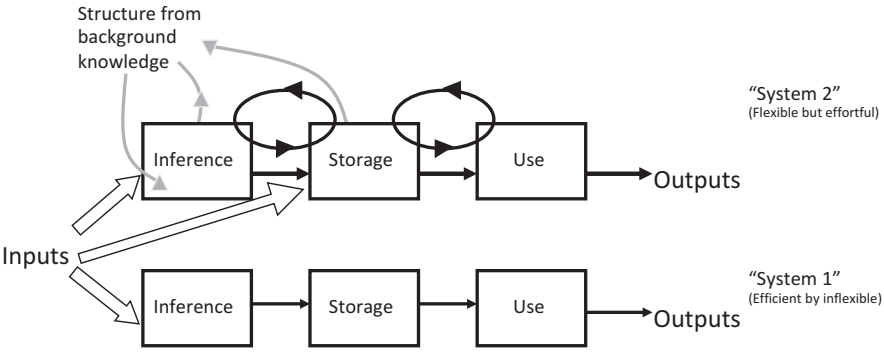

Fig. 1 A "two systems" model of mindreading (simplified from Apperly, 2010). The model distinguishes processes involved in inference, storage and use of information about others' mental states. "System 2" makes flexible, context-sensitive mindreading inferences by drawing richly upon background knowledge, in processes represented by the grey arrows. Oval arrows indicate that System 2 mindreading will often involve repeated cycles of reasoning. System 1 processes manage to be more cognitively efficient by limiting their interaction with background information and limiting their processing over inputs. For clarity, only one System 1 process is depicted, but there are likely to be multiple processes, for example to enable mindreading of belief-like states, goals and emotions

Samson et al., 2010; van der Wel et al., 2014), while reduced flexibility is evidenced by the appearance of automaticity only for relatively simple problems (such as inferring what someone sees, Samson et al., 2010) and not for more complex problems (such as inferring precisely how they see it from their perspective; Surtees, Samson, & Apperly, 2016). The greater number and complexity of arrows for System 2 reflects the greater flexibility of information flow compared with System 1. The main focus of the present chapter will be on System 2 processes.

Implicit in Fig. 1 is the fact that mindreading requires the representation of someone else and their mental states as distinct from one's self and one's own. Maintaining this distinction is essential, but it is also challenging because the perspective of the other and of one's self are not independent records of facts, but are related to each other and to "reality". This closely related information gives rise to interference between self and other, such that if I represent our differing beliefs about something (even something as mundane as the location of a hidden object), I am slower and more error-prone when judging what you think ("egocentric interference", Royzman, Cassidy, & Baron, 2003), and when judging what I think myself ("altercentric interference", Samson et al., 2010). A successful mindreader must not only maintain a distinction between the perspectives of self and other, but also manage the interference that results: mindreading requires inference, representation and control.

Figure 1 helps systematize a set of important questions about mindreading. For example, are any or all of these processes specialized for mindreading; do the cognitive control requirements arise at all stages of processing; is the network of brain regions implicated in mindreading equally involved in inference, storage and use of mindreading information? In the next section I will tackle some of these questions, and show how a cognitive model helps us understand what light cognitive neuroscience has already shone on our understanding of mindreading.

Specialization for Mindreading

While mindreading does not appear to make unique reasoning demands, a related hypothesis is that the cognitive and neural systems for mindreading are domain-specific. The latter does not entail the former, because reasons other than unique reasoning demands could lead mindreading to show domain specificity. For example, if there is neural specialization for other social processes (e.g. Adolphs, 2009; Frith, 2007) neural activity during mindreading may show domain specificity for at least three reasons: (1) because one or more of those other social processes are intrinsic mindreading, (2) because those social processes have distinctive neural connectivity with neural systems involved in mindreading, (3) because mature mindreading develops on the foundation of other social processes that are themselves domain-specific, and so inherits domain specificity without this being functionally necessary.

Domain specificity for mindreading has been tested most extensively in a series of studies by Saxe and colleagues (Koster-Hale & Saxe, 2013; Saxe & Kanwisher,

2003). This widely adopted approach starts by contrasting neural activation while participants reason about false beliefs with activation during structurally and logically similar reasoning about false photographs and false signs. Brain activity surviving this contrast is then tested for its selectivity for a range of other judgements about people's mental states, personality, physical appearance and other characteristics. While the contrast between false beliefs and false photos typically reveals activity in mPFC, bilateral TPJ and temporal poles, over an impressive range of studies it is right TPJ that shows the highest selectivity for reasoning about mental states (Koster-Hale & Saxe, 2013).

These results illustrate the value of cognitive neuroimaging for understanding mindreading because they provide stronger evidence than behavioural studies that mindreading involves domain-specific processes. However, there are also important caveats. First, demonstrating domain specificity is just one step towards understanding underlying mechanisms, and for now it remains unclear what function rTPJ is performing or what feature of mindreading leads to evidence of domain specificity (see Future Prospect, below). It is not clear whether domain-specific processes are involved in inference, storage or use of mindreading information, or all three (Fig. 1). Second, as described above, it is clear that mindreading depends upon many processes, which will not all be domain-specific. It's therefore important that questions about domain specificity are complemented by questions about the broader functional basis of mindreading. Third, the best methods for testing the domain specificity of mindreading are unsuitable for understanding these broader components of mindreading because such processes are subtracted out of the comparison between strictly matched mindreading and non-mindreading tasks. The most obvious examples of this are processes involved in the control of mindreading.

Control Processes During Mindreading

Control of egocentrism A vivid illustration that domain-general processes contribute significantly to mindreading comes from the neuropsychological case study of patient WBA (Samson, Apperly, Kathirgamanathan, & Humphreys, 2005). This patient sustained a right frontal brain lesion, following a stroke, and his lesion affected lateral frontal brain regions most commonly implicated in cognitive control, notably including right inferior frontal gyrus. Consistent with this WBA showed notable impairment on standard neuropsychological assessments of executive function, including inhibitory control. Medial PFC—commonly implicated in mindreading—was left largely intact. Consistent with this, WBA appeared to be able to reason about other people's false beliefs, provided he was tested on an unusual task that minimized the salience of his own knowledge of the correct answer. However, on more standard false belief tasks and on a range of other tests of his ability to judge other people's perspectives he showed very high rates of "egocentric errors", where he responded according to his own perspective rather than the

other person's. Anecdotal report from a family member indicated that this pronounced egocentrism was not limited to laboratory tasks.

Importantly, such egocentric errors are not simply the product of generic task difficulty. In a follow-up study WBA, and another patient with similar brain injury, showed egocentric errors when required to judge the differing desires of an opponent in a card game, but lower errors when judging the card they next needed themselves, despite variation in whether a matching or a mismatching card would be a winner. A second pair of patients with lesions to more medial prefrontal cortex showed the opposite pattern of errors (Samson, Houthuys, & Humphreys, 2015). This demonstrates a classical neuropsychological double dissociation between the control processes necessary for managing interference from self perspective when taking the other's perspective, versus those necessary for handling conflict arising from other aspects of game strategy.

Such evidence from studies of patients converges with evidence from fMRI, ERP and TMS in suggesting a selective role for lateral frontal regions—in particular inferior frontal gyrus—in controlling tendencies for both egocentric and altercentric error and bias during mindreading (e.g. McCleery, Surtees, Graham, Richards, & Apperly, 2011; van der Meer et al., 2011; Vogeley et al., 2001). For example, Hartwright, Apperly, and Hansen (2012) used a "belief-desire" task in which participants used a character's beliefs and desires to predict their search in one of two boxes. Participants were told which box contained some food, which box the character thought contained the food, and whether or not the character desired the food on that trial. When the character's belief was false there was conflict between his perspective and the participants', but not when his belief was true. In contrast the character's desire for the food was not systematically related to the participants' (he might like peas, whereas the participant does not), so conflict was equally likely to occur (or not occur) at each level of this factor. Consistent with previous behavioural studies (e.g. Apperly, Warren, Andrews, Grant, & Todd, 2011; German & Hehman, 2006), responses were slower and more error-prone whenever the character's belief was false and whenever his desire was negative. A natural interpretation of these results might be that the belief and desire effects were equivalent, perhaps because false belief and negative desire both required more inhibitory control (Friedman & Leslie, 2004). However, fMRI data suggested that these effects were not equivalent: whereas activity in bilateral TPJ and dorsomedial PFC was influenced by both belief and desire, activity in right IFG was influenced only by the factor of belief, and not by the factor of desire. Moreover, in a subsequent study, r-TMS to right IFG influenced performance on false versus true belief trials, and not negative versus positive desire trials (Hartwright, Hardwick, Apperly, & Hansen, 2016). These findings converge with the neuropsychological evidence in suggesting that IFG is involved specifically in resisting "egocentric" interference from self perspective when taking the perspective of someone else.

Self versus other The need to control interference from self perspective when taking the perspective of another presupposes that you have represented the other's perspective. In parallel with work on controlling egocentrism is a burgeoning

literature on the cognitive and neural basis of distinguishing self from other (e.g. Cook, 2014). This work began with evidence that observing another's action creates a tendency for "automatic imitation" of the action by one's self, which must be controlled if a different action is necessary for the task (Brass, Bekkering, Wohlschläger, & Prinz, 2000). Whereas controlling interference from other kinds of over-learned association is typically linked with activity in lateral prefrontal brain regions (e.g. Wagner, Maril, Bjork, & Schacter, 2001), control of automatic imitation appears to depend on regions of mPFC and TPJ similar or identical to those commonly implicated in mindreading. A number of studies suggest that this link with mindreading is more than coincidental. For example, Santiesteban, White, et al. (2012) found that training inhibition of automatic imitation improved participants' use of mindreading in a communication task (the Director Task; Keysar, Lin & Barr, 2003), whereas training generic inhibition did not. Santiesteban, Banissy, Catmur, and Bird (2012) found that stimulation of rTPJ improved both inhibition of imitation and use of mindreading in a communication task. Such findings suggest that the same process of self-other control may be at work in both imitation inhibition and perspective-taking, with one hypothesis being that TPJ maintains the distinction between information related to self versus other (perhaps in line with its role in general control of attention), while mPFC prioritizes one or other set of information according to the task or the context (Santiesteban, Banissy, et al., 2012).

An assimilation On the face of it, these data appear contradictory to those presented in the previous section: "self/other control" and "control of egocentrism" sound a lot like two terms for the same phenomenon, yet the data suggest they depend on different functional and neural processes. I suggest, however, that if we think about mindreading in terms of component processes then there may be no contradiction. An example will help illustrate the point. McCleery et al. (2011) used a simple perspective-taking task in which participants viewed a schematic room with dots on the wall and an avatar standing in the middle. The avatar's position meant that the number of dots he saw was sometimes consistent with the participant's perspective and sometimes inconsistent. On some trials participants were told to judge how many dots they themselves saw when the picture appeared (self trials) while on other trials they judged how many the avatar saw (other trials). Participants are slower to judge both self and other perspectives whenever those perspectives are inconsistent (Samson et al., 2010), and a simultaneous executive task increases this effect to an equal degree for self and other judgements (Qureshi et al., 2010). We have interpreted this pattern to suggest that self and other perspectives are calculated on every trial in a relatively effortless manner ("inference" in Fig. 1), with the effortful step being a subsequent process of selecting either self or other perspective as the basis for a response ("use" in Fig. 1). McCleery et al. (2011) recorded ERPs during this task. They observed a component from electrodes over temporoparietal cortex approximately 450 ms after picture onset, which varied according to whether participants were making self or other judgements. They also observed a later and longer-lasting component from electrodes over right frontal cortex, which varied only according to whether self and other perspectives were consistent versus

inconsistent. These effects were tentatively localized to left and right TPJ and right IFG, respectively.

I suggest that these results support generalizable conclusions that help make sense of a variety of findings about control processes during mindreading. Mindreading requires the establishment and maintenance of a distinction between self and other, which depends on TPJ. This may well be necessary at all processing steps: inferences, storage and use (Fig. 1). Having distinguished self and other we are then in a position to use either self or other perspective to make responses or to inform further processing. Whichever perspective we are trying to use, the other perspective will tend to compete, potentially activating the response relevant to the opposite perspective from the one intended. This latter interference may originate with representations of perspectives but in other respects resembles entirely generic interference effects and recruits generic processes associated with IFG. It occurs most clearly during the use of mindreading information (Fig. 1), and difficulty with resisting this interference leads to a large number of the egocentric phenomena reported in the literature.

Mindreading Inferences

While almost all mindreading tasks require participants to infer a target's mental states, Fig. 1 encourages us to distinguish such inferences from other mindreading processes. And just as Tolstoy can tell us what Anna is thinking, and real people can inform us of their thoughts and feelings, so we can create experimental tasks that remove the need to infer the mental states of others. Among other things, this allows us to ask whether any brain areas involved in mindreading are distinctively involved in such mindreading inferences. The belief-desire task described earlier (Apperly et al., 2011; Hartwright et al., 2012) opens this possibility, because participants are simply told the character's belief and desire. In terms of Fig. 1, participants skip the initial inference step, but must *store* the mental states they are told and *use* them to reason about the character's behaviour. Hartwright et al. also employed the false belief/false photograph "localizer" task developed by Saxe and colleagues, which clearly does involve mindreading inferences. In the belief-desire task, variation in the character's belief and desire modulated activity in bilateral TPJ, showing substantial overlap with TPJ voxels identified in the false belief/false photograph task. In contrast, neither the belief nor the desire factor modulated activity in ventral mPFC, though this brain region did show selective activity in the false belief/false photograph task. Participants in this study were clearly capable of engaging v-mPFC for mindreading, but did not appear to do so when they only had to store and use mental states to predict behaviour. In a second study, Hartwright et al. (2014) adapted the belief-desire task to reintroduce the need for a mindreading inference. In this task the character changed from trial to trial, there were prizes rather than foods, and the character's desire for the prize on offer was indicated through realistic

photographs of faces that were smiling (positive desire), frowning (negative desire), or neutral (unknown desire). In the unknown desire condition participants had to make a mindreading inference about whether that character would want that prize. In this task, variation in the desire factor did modulate activity in v-mPFC, and this effect was driven by the unknown desire condition differing from the positive and negative desire conditions. These findings suggest that v-mPFC may have a distinctive role in mindreading inferences, and that the near ubiquity of activity in this region in studies of mindreading reflects the fact that most mindreading tasks entail mindreading inferences, and cannot distinguish activity due to these inferences from other component processes.

Of course, associating mindreading inferences with v-mPFC is just one step in understanding the cognitive basis of mindreading inferences and what role v-mPFC has in supporting these processes. As discussed earlier, mindreading inferences often involve complex integration of information from multiple sources under conditions of uncertainty in order to make a "best guess" about the target's mental states. The apparent simplicity of classic mindreading tasks, such as the "Sally-Anne" task, obscures the fact that it is only the pragrmatic context that suggests she must think it's either in the basket or the box: in fact Sally *could* think her ball is absolutely anywhere. Such uncertainty and context-sensitivity is much more apparent in more realistic mindreading situations (Apperly, 2010). The hypothesis that v-mPFC helps meet these functional requirements is supported by a study from Jenkins and Mitchell (2010) who independently varied whether a mindreading task required inferences about a character's mental states or their preferences, and whether those inferences were clearly warranted by the situation or were more uncertain and ambiguous. Whereas TPJ (and not mPFC) activity was sensitive to whether the inferences concerned mental states versus preferences, mPFC activity (and not TPJ) was sensitive to the level of uncertainty in the inference. Moreover, these findings converge with a broader literature that implicates v-mPFC in complex information integration and reasoning under uncertainty (e.g. Burgess, Dumontheil, & Gilbert, 2007).

A Future Prospect: Do Mindreading Brain Regions Represent What Others Are Thinking?

Since mindreading involves representing what other people are thinking (or feeling, or intending, etc.), and since mindreading recruits a reliable network of brain areas, it would be natural to suppose that one or all of these brain areas represents the thoughts of other people. Surprisingly, however, no evidence bears directly on this question, and in fact different theories about the "mindreading brain network" point towards different expectations. It is exciting that methods for decoding the informational content of neural activity (e.g. Haxby, Connolly, & Guntupalli, 2014; Kriegeskorte & Kievit, 2013) are opening up the possibility of directly testing such questions.

The idea that the "mindreading brain network" must be representing what other people are thinking seems a good hypothesis, and it clearly predicts that during a mindreading task TPJ and/or mPFC must be carrying information that distinguishes between instances in which Sally thinks her marble is in the basket, versus Sally thinks her marble is in the box, versus John thinks his marble is in the basket, etc. Put more operationally, if one trained a multivariate pattern classifier on patterns of activity in TPJ (for example) over a variety of instances in which an agent thinks an object is in a location, the classifier should be able to take new data from the same subject and distinguish trials on which Sally thinks the ball is in the basket from other combinations of information about agent-object-location combinations. Encouragingly, recent evidence suggests that category-level and even item-level information can be decoded from patterns of activity in TPJ and mPFC during memory retrieval (Kuhl & Chun, 2014; Zeithamova, Dominick, & Preston, 2012), suggesting that this question is tractable for suitably designed sets of Agent-Object-Location stimuli.

However, this outcome is far from a foregone conclusion. It is well-known that TPJ and mPFC are involved in attentional control, as well as mindreading (e.g. Burgess et al., 2007; Corbetta, Patel, & Schulan, 2008), and as discussed earlier there are good grounds for thinking that TPJ and mPFC may be specifically involved in controlling attention in order to maintain a distinction between information and processes related to self and other. This is compatible with the selective engagement of TPJ and mPFC in mindreading, but in no way entails that these regions represent the information about the agents, objects, locations, etc. over which they are exerting control; instead that information could be represented in participants' own primary semantic systems. Thus we do not yet have an answer to one of the most fundamental neuroscientific questions about mindreading: do mindreading brain regions represent information about mindreading?

Studies of mindreading have just begun to exploit the power of MVPA, successfully decoding broad types of social tasks and mental states from activation patterns in TPJ and/or mPFC (e.g. Koster-Hale, Saxe, Dungan, & Young, 2013; Tamir, Thornton, Contreras, & Mitchell, 2016). Extending this approach to examine how and when we represent the content of other minds not only addresses questions about how the brain supports mindreading. It also opens ways to tackle functional questions that have proved fiendishly difficult to address so far: Do perspectives of self and other recruit the same representational resources? Are self and other perspectives activated in series or in parallel? Do control processes, such as those associated with IFG, work to resolve competition between the content of self and other perspectives, or only competition between responses or judgements based on these perspectives. The role of IFG in inhibiting representational content during selective episodic memory retrieval (e.g. Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015) certainly makes it plausible that IFG also directly acts on the contents of self and other perspectives. In sum, MVPA offers the prospect of a rich interaction between cognitive and neuroscientific approaches through the common currency of "information".

Summary I have outlined a cognitive model of mindreading that is narrowly focused on processes directly involved in inferring, storing and using information about other people's mental states. A narrow focus makes it possible to think about the relationships between individual processing steps and their cognitive and neural bases, but of course it should not blind us to the fact that there is much more to mindreading than what I have discussed here. More ambitious and exhaustive models are very valuable but they face a daunting challenge in knowing where to stop. A good case can be made for including gaze processing, face recognition, moral and causal reasoning as part of mindreading (e.g. Schaafsma et al., 2015), However, following this logic, since I can imagine you thinking anything I can think for myself, there seems no principled limit on the information and processes on which I might need to draw, and so no straight-forward way of distinguishing between processes that are involved and not involved in mindreading. This is a deep issue with mindreading, but it should not stop us from building rich models of how mindreading is supported by a variety of cognitive and neural processes.

I hope I have also demonstrated that this is a two-way street, with results from neuroscientific studies informing cognitive theories just as much as the reverse. Relevant theories and methods must also interact. For example, it is important to recognize that subtractive neuroimaging designs optimized to detect domain-specific mindreading processes will tell us little about the nature of the processes involved, whereas designs that contrast different conditions within a mindreading task might tell you more about processes but little about their domain specificity. The rate of innovation in neuroscientific methods holds out great future promise for a cognitive neuroscience of mindreading, which will be maximized when combined with functional models of the cognitive processes involved.

References

Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology, 60*, 693–716.

Apperly, I. A. (2008). Beyond simulation-theory and theory-theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind". *Cognition, 107*, 266–283.

Apperly, I. A. (2010). *Mindreaders: The cognitive basis of "theory of mind"*. Hove, UK: Psychology Press/Taylor & Francis Group.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970.

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science, 17*(10), 841–844.

Apperly, I. A., Warren, F., Andrews, B. J., Grant, J., & Todd, S. (2011). Error patterns in the belief-desire reasoning of 3- to 5-year-olds recur in reaction times from 6 years to adulthood: Evidence for developmental continuity in theory of mind. *Child Development, 82*(5), 1691–1703.

Barwise, J., & Perry, J. (1983). *Situations and attitudes*. Cambridge, MA: MIT Press.

Brass, M., Bekkering, H., Wohlschläger, A., & Prinz, W. (2000). Compatibility between observed and executed finger movements: Comparing symbolic, spatial, and imitative cues. *Brain and Cognition, 44*(2), 124–143.

Burgess, P. W., Dumontheil, I., & Gilbert, S. J. (2007). The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences, 11*(7), 290–298.

Cook, J. L. (2014). Task-relevance dependent gradients in medial prefrontal and temporoparietal cortices suggest solutions to paradoxes concerning self/other control. *Neuroscience and Biobehavioral Reviews, 42*, 298–302.

Corbetta, M., Patel, G., & Schulan, G. L. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron, 58*, 306–324.

Davidson, D. (1990). The structure and content of truth. *Journal of Philosophy, 87*(6), 279–328.

Davies, M., & Stone, T. (Eds.). (1995). *Folk psychology: The theory of mind debate*. Oxford, England: Blackwell.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Fauconnier, G. (1985). *Mental spaces: Aspects of meaning construction in natural language*. Cambridge, MA: MIT Press.

Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Friedman, O., & Leslie, A. M. (2004). Mechanisms of belief-desire reasoning: Inhibition and bias. *Psychological Science, 15*, 547–552.

Frith, C. D. (2007). The social brain? *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*(1480), 671–678.

German, T., & Hehman, J. (2006). Representational and executive selection resources in 'theory of mind': Evidence from compromised belief-desire reasoning in old age. *Cognition, 101*(1), 129–152.

Hartwright, C. E., Apperly, I. A., & Hansen, P. C. (2012). Multiple roles for executive control in belief-desire reasoning: Distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. *NeuroImage, 61*(4), 921–930.

Hartwright, C., Apperly, I. A., & Hansen, P. C. (2014). Representation, Control or Reasoning? Distinct Functions for Theory of Mind within the Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience, 26*(4), 683–698.

Hartwright, C. E., Hardwick, R., Apperly, I. A., & Hansen, P. (2016). Structural morphology in resting state networks predict the effect of theta burst stimulation in false belief reasoning. *Human Brain Mapping, 37*, 3502–3514.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience, 37*, 435–456.

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science, 9*, 131–143.

Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking.* Cambridge, MA: Harvard University Press.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences, 20*(8), 589–604.

Jenkins, A. C., & Mitchell, J. P. (2010). Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex, 20*(2), 404–410.

Keysar, B., Lin, S., Barr, D.J., (2003). Limits on theory of mind use in adults. *Cognition 89*, 25–41.

Koster-Hale, J., & Saxe, R. (2013). Functional neuroimaging of theory of mind. In S. Baron-Cohen, M. Lombardo, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (pp. 132–163). Oxford, England: Oxford University Press.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Science, 110*, 5648–5653.

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.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*, 401–412.

Kuhl, B. A., & Chun, M. M. (2014). Successful remembering elicits event-specific activity patterns in lateral parietal cortex. *The Journal of Neuroscience, 34*(23), 8051–8060.

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

Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences, 9*(10), 459–462.

Low, J., Apperly, I. A., Butterfill, S. A., & Rakoczy, H. (2016). Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account. *Child Development Perspectives, 10*(3), 184–189.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.

McCleery, J. P., Surtees, A. D., Graham, K. A., Richards, J. E., & Apperly, I. A. (2011). The neural and cognitive time course of theory of mind. *The Journal of Neuroscience, 31*(36), 12849–12854.

Perner, J., & Leekam, S. (2008). The curious incident of the photo that was accused of being false: Issues of domain specificity in development, autism, and brain imaging. *The Quarterly Journal of Experimental Psychology, 61*(1), 76–89.

Phillips, J., Ong, D. C., Surtees, A. D., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science, 26*(9), 1353–1367.

Qureshi, A., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective-selection, not Level-1 visual perspective-calculation: Evidence from a dual-task study of adults. *Cognition, 117*(2), 230–236.

Royzman, E. B., Cassidy, K. W., & Baron, J. (2003). "I know, you know": Epistemic egocentrism in children and adults. *Review of General Psychology, 7*(1), 38–65.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (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.

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: A case of selective deficit in inhibiting self-perspective. *Brain, 128*, 1102–1111.

Samson, D., Houthuys, S., & Humphreys, G. W. (2015). Self-perspective inhibition deficits cannot be explained by general executive control difficulties. *Cortex, 70*, 189–201.

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Current Biology, 22*, 2274–2277.

Santiesteban, I., White, S., Cook, J., Gilbert, S. J., Heyes, C., & Bird, G. (2012). Training social cognition: From imitation to theory of mind. *Cognition, 122*, 228–235.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage, 19*, 1835–1842.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19*(2), 65–72.

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience, 10*, 640–656.

Stuhlmüller, A., & Goodman, N. D. (2014). Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research, 28*, 80–99.

Surtees, A., Samson, D., & Apperly, I. A. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition, 146*, 97–105.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Science, USA, 113*(1), 194–199.

van der Meer, L., Groenewold, N.A., Nolen, W.A., Pijnenborg, M., Aleman, A., (2011). Inhibit yourself and understand the other: Neural basis of distinct processes underlying Theory of Mind. *NeuroImage 56*, 2364–2374.

van der Wel, R. P., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition, 130*(1), 128–133.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., … Zilles, K. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage, 14*(1).

Wagner, A. D., Maril, A., Bjork, R. A., & Schacter, D. L. (2001). Prefrontal contributions to executive control: fMRI evidence for functional distinctions within lateral prefrontal cortex. *Neuroimage, 14*(6), 1337–1347.

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories. *Nature Neuroscience, 18*, 582–589.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75*, 168–179.