Challenging the Two-systems Model of Mindreading

Pierre Jacob

5

Challenging the Two-systems Model of Mindreading

Pierre Jacob

1. Introduction

This chapter is devoted to the two-systems model of mindreading recently advocated by the psychologist Ian Apperly and the philosopher Steve Butterfill. Mindreading or theory-of-mind is the human social cognitive ability to represent the contents and attitudes of the psychological states of either self or others. Philosophers have addressed the topic of mindreading or theory-of-mind for several different special purposes. Theory-of-mind (or folk psychology) has been at the centre of *ontological* controversies over the mind–body problem about the fundamental nature of mental states.[1] In the context of responding to the challenges of scepticism, mindreading has also been central to *epistemological* discussions of how one can know that other people have mental states (the problem of other minds), how one can know the contents of one's own mind (the problem of introspection), and how both kinds of knowledge are related to knowledge of the non-mental world.[2]

Psychologists and cognitive scientists investigate the psychological mechanisms involved in mindreading. To attribute a psychological state to either oneself or another individual is to form a belief (or judgement) about the content and the attitude of one's own or another's psychological state. To the extent that an ascribed (or represented) psychological state can itself be construed as a mental representation, a mindreader's belief or judgement about the content of her own or another's psychological state can in turn be construed as a mental representation of a mental representation, i.e. as a *meta-representation*. Thus, Leslie (1987, 1988), Sperber (1985, 2000) and others have argued that the human mindreading ability is best construed as a meta-representational capacity whereby one's own system of internal mental representations can serve as its own meta-language (Sperber 2000).

Full-blown mindreading is often taken to be an effortful cognitive capacity on the grounds that it rests on meta-representational resources. But on reflection, this assumption may turn out to be a prejudice (see section two). Advocates of

[1] Cf. Stich and Nichols (2003).
[2] For a particularly interesting example, cf. Davidson (1991a).

Pierre Jacob, *Challenging the Two-systems Model of Mindreading*. In: *Knowing Other Minds*. Edited by Anita Avramides and Matthew Parrott, Oxford University Press (2019).
 DOI: 10.1093/oso/9780198794400.003.0005

the two-systems model argue for the existence of a *minimal*, fast, and efficient mindreading capacity distinct from full-blown mindreading in that it falls short of being meta-representational. This model is interesting and challenging. However, my goal in this chapter is not to praise it but to appraise and even to challenge it. The chapter is divided into eight sections (including the introduction and the conclusion). As I explain in section 2, the main purpose of the two-systems model of mindreading is to resolve the cognitive tension between efficiency and flexibility. In section 3, I spell out the basic contrast between the contents of beliefs and registrations: while the representation of the former is supposed to be effortful and to require full-blown mindreading, the representation of the latter is supposed to be achievable by the minimal mindreading system. In section 4, I describe the fundamental developmental puzzle, whose resolution is one of the main rationales for the two-systems model. In section 5, I assess the attempted resolution of this puzzle by the two-systems model. In section 6, I address the question whether the aspectuality of beliefs is a 'signature limit' of the minimal mindreading system. In section 7, I examine the contrast between the putative automaticity of Level-1 visual perspective-taking tasks (which are allegedly performed by the minimal system) and the effortful resolution of Level-2 visual perspective-taking tasks (which allegedly require the flexibility of the full-blown mindreading system).

2. Seeking A Middle Ground

Ever since Premack and Woodruff's (1978) paper, much cognitive scientific and psychological investigation of mindreading of the past forty years or so has been devoted to four related basic empirical questions:

1) To what extent does the meta-representational architecture of mindreading rest on the resources of the human language faculty?
2) To what extent is mindreading unique to humans?
3) How ubiquitous is mindreading in human adult social cognition?
4) To what extent do human children learn to read minds through a cultural process involving language acquisition?

There is presently a lively and unresolved controversy over these four related empirical questions between advocates of what can be called a 'cultural constructivist' approach to mindreading and their nativist critics. The former assume and the latter deny that, while the meta-representational architecture of mindreading rests on the human language faculty, mindreading is unique to human adults,[3]

[3] In their (2008) review, Call and Tomasello argued that false-belief understanding is unique to humans. But the recent *Science* paper by Krupenye et al. (2016) provides evidence for false-belief understanding in a variety of great apes.

it is not ubiquitous in human adult social cognition, and human children learn their mindreading skills through a cultural process involving language acquisition and linguistic transmission.[4]

As Perner and Ruffman (2005: 214) have put it on behalf of the cultural constructivist approach, the human mindreading ability 'may be constructed in a cultural process tied to language acquisition'. Some philosophical advocates of a so-called 'radical enactivist' approach to human social cognition have further argued that non-human animals and prelinguistic human infants are likely to lack the meta-representational resources necessary for full-blown mindreading on the grounds that they can merely entertain what Hutto (2008) calls 'intentional attitudes', not propositional attitudes, i.e. mental representations with genuine contents. If this were true, then non-human animals and preverbal infants could not, nor perhaps would they need to, use the contents of their own propositional attitudes in order to meta-represent the contents of others' propositional attitudes. Other philosophers have argued that much human social cognition rests on what they call 'second-person primary interactions' which they take to be independent from mindreading capacities (Gallagher 2001). Other cognitive scientists have argued that the evidence for mindreading in non-human primates and prelinguistic infants is best construed as a capacity for either behaviour-reading (cf. Penn and Povinelli 2007) or for sub-mentalizing (cf. Heyes 2014).[5]

In a nutshell, advocates of cultural constructivism find it incredible that preverbal infants may have hardwired meta-representational resources; they assume that human children acquire their mindreading abilities through language acquisition. From the standpoint of their critics, cultural constructivist approaches to mindreading face two related challenges, the first of which has been put forward by Sperber (2000: 120) thus: if 'an organism endowed with a rich internal system of conceptual representations' lacked the ability 'to use these "opaquely" or meta-representationally, that is, as iconic representations of other representations', then the question would arise how she could *learn* to do so on the basis on her ontogenetic developmental experience. The second related challenge is that unless they could read their caretakers' minds, it is quite unclear how human infants could learn their native tongue.

Recently, the question has arisen whether there could be a social cognitive mechanism that is the *middle ground* between full-blown meta-representational mindreading and either behaviour-reading or sub-mentalizing. According to the advocates of the two-systems model of mindreading, there is room for such a

[4] Heyes and Frith (2014) have recently argued that human children learn to read others' minds in the same way they learn to read words. Cf. Strickland and Jacob (2015) for a critical discussion.

[5] According to Heyes's sub-mentalizing approach, humans can solve what seems like mindreading tasks by employing 'domain-general cognitive processes that do not involve thinking about mental states but can produce in social contexts behaviour that looks as if it is controlled by thinking about mental states' (2014: 132). For a detailed criticism, see Jacob (2018).

middle ground, which they call *minimal* mindreading (or *minimal* theory-of-mind) and which is identical with neither full-blown mindreading nor behaviour-reading, let alone sub-mentalizing (cf. Apperly 2011; Apperly and Butterfill 2009; Butterfill and Apperly 2013; Low et al. 2016).

Advocates of the two-systems model take full-blown mindreading to be a uniquely human meta-representational cognitive capacity. They further take it to be effortful (or costly), normative and flexible, to rest on language, and to emerge slowly in human ontogenetic development. Finally, they endorse the view that full-blown theory-of-mind enables one to represent not just others' (or one's own) psychological states but also one's own and others' *reasons*. This is what Apperly (2011) and Apperly and Butterfill (2009) call 'the *normative* account of mindreading'. By contrast, minimal (i.e. non-meta-representational and non-normative) mindreading is taken to be fast, inflexible, efficient, and non-normative: it is taken to emerge earlier than full-blown mindreading in human ontogenetic development, not to depend on language, nor to be uniquely human. Crucially, minimal (or early-developing) mindreading is *not* supposed to grow into, nor to be superseded by, full-blown mindreading. The two systems are separate and do not speak to each other: the early-developing, fast, and efficient system is supposed to persist throughout development into adulthood alongside the later developing system. As a result, one should be able to find evidence of the 'signature limits' of the early-developing system in adulthood.

This two-systems model of mindreading is one among several versions of two-systems models of human cognitive architecture that have emerged in recent cognitive science. For example, the two-systems model of human vision rests on the discovery of basic dissociations between visual perception and visually guided actions (Goodale and Milner 1995; Jacob and Jeannerod 2003). Dual models of human reasoning rest on the distinction between intuitive and reflective responses elicited by logical problems (Kahneman 2003, 2011). Recent work on numerical cognition suggests a dichotomy between core numerical systems and the language-based full-blown capacity to represent integers (Feigenson et al. 2004). More directly related to the study of human social cognition, in the context of his investigation of the role of mental simulation in tasks of mindreading, Goldman (2006) has drawn a distinction between low-level and high-level processes of simulation that reflects the distinction between mirroring (or the activity of mirror neurons) and the imagination. Each of these two-systems models of human cognitive architecture must be judged on its own merits.

What lies at the core of Apperly and Butterfill's two-systems model of mindreading is the recognition that mindreading is subject to the *conflicting* cognitive demands of *flexibility* and *efficiency*: while a soccer player needs a fast and efficient system that will enable him to deceive a goalkeeper in a split second, a jury or a judge needs a flexible but effortful system that will enable her to reflect over several days, if not months, about a defendant's motivations and epistemic states

Minimal mindreading	*Full-blown mindreading*
• early-developing	• later-developing
• implicit	• explicit
• non-normative	• normative
• fast	• slower
• efficient	• effortful
• automatic	• controlled
• inflexible	• flexible
• encapsulated	• unencapsulated

Figure 5.1 Apperly and Butterfill (2009)

(Apperly 2011; Low and Watts 2013). Only a cognitive mechanism capable of meta-representing the contents of the defendant's mental states and of representing his reasons for his actions could achieve what a judge needs. As Apperly (2011: 9) puts it,

> the difficulty in having one system that is both flexible and efficient is apparent from the high prevalence of 'two-systems' accounts in cognition, whereby in a given domain…the contradiction is resolved by having two types of cognitive system that operate in the domain, which make complementary trade-offs between flexibility and efficiency.

As Apperly and Butterfill (2009: 264) earlier put it, 'our central claim is that early-developing and late-developing systems for belief processing need to make different and complementary trade-offs between flexibility and efficiency' (cf. Figure 5.1).

3. Why Full-Blown Mindreading Is Taken to Be Effortful

Human mindreading underlies the ascription to others of a wide variety of psychological states, including emotions (and other affective states), motivations (e.g., desires and intentions), and epistemic states (perceptions, beliefs, states of knowledge). Much of the experimental developmental investigation of mindreading of the past thirty years or so rests on the fundamental and widespread assumption, since the publication and the discussion of Premack and Woodruff's (1978) paper, that *false-belief understanding* is a decisive mark of mindreading. The capacity for false-belief attribution is widely taken to demonstrate one's understanding that an agent's (instrumental) action does not depend merely on non-mental features of her environment, but on her mental representation of her environment. This is one of the major reasons why advocates of the two-systems model focus on the representation of the contents of others' *epistemic* states.

As Butterfill and Apperly insightfully argue, one might track toxicity, not by representing it *as such*, but instead by representing another property that is reliably correlated with toxicity, e.g., foul odors. As they focus on the representation of others' epistemic states, the main relevant question for them is: 'What could someone represent that would enable her to track, at least within limits, others' perceptions, knowledge states and beliefs including false beliefs?' (2014: 606). What they call 'minimal theory-of-mind' involves the representation of belief-like states, 'but it does not involve representing beliefs or other propositional attitudes *as such*' (2014: 607). However, as they are well aware, it is not likely that one could track the contents of *all* of another's beliefs (true or false), including e.g., the contents of others' beliefs about object-identity, *without* representing them *as such* (this is why their penultimate quote in this paragraph contains the clause 'at least within limits'). Advocates of minimal theory-of-mind call such belief-like states *registrations*. They argue that by representing the contents of others' registrations, one can track the contents of a restricted range of others' beliefs, namely beliefs about an object's location. So the question naturally arises: what makes the contents of registrations really different from the contents of genuine beliefs about an object's location?

There are four related reasons why representing the contents of others' beliefs (and propositional attitudes) *as such* has been taken to be an effortful psychological process by advocates of the two-systems model and others: the meta-representational architecture of full-blown mindreading; the role of reasons in explaining, evaluating, and justifying human action; confirmation holism; and the aspectuality of propositional attitudes.

I start with the meta-representational architecture of full-blown mindreading. On our best current understanding, full-blown mindreading is the meta-representational (or meta-psychological) capacity to form a higher-order representation of a lower-order representation where the latter is embedded within the former as in (1) (where Ann is attributing a belief to Marta):

(1) Marta believes that the evening star is shining.

The purpose of the higher-order representation is to attribute to an agent (Marta) the lower-order representation. Whether or not Ann endorses the belief that the evening star is shining, she must be able to entertain the thought that the evening star is shining in order to ascribe to Marta the belief that the evening star is shining. The fact that our best current understanding of full-blown mindreading exhibits a complex meta-representational architecture need not entail that the use of this capacity by human adults is demanding or effortful. Nor does the fact that our best current scientific characterization of the primate visual system exhibits a complex computational architecture entail that visual processing is particularly effortful to primates.

I now turn to the role of reasons in the explanation and justification of human actions. In their paper devoted to the possibility of mindreading in non-human

primates, Premack and Woodruff (1978) identified mindreading (i.e. the imputation of mental states to others) with the possession of a *theory-of-mind* on the joint grounds that the imputed mental states are *unobservable* and that the imputation underlies the *prediction* of others' behaviour.

As philosophers of science have emphasized, predicting is not explaining. It is controversial to some extent whether possession of a theory-of-mind (i.e., the capacity to attribute psychological states to others) is necessary for predicting an agent's instrumental action in all cases (cf. Andrews 2003; Perner and Roessler 2010). However, it is much less controversial that the attribution of psychological states to others is necessary for both *explaining* and *evaluating* the success or failure of actions (Andrews 2009). Both the explanation and the normative evaluation of an agent's instrumental action rest on one's ability to represent her *reasons* for her actions. This has opened the path for Davidson's (1970) well-known argument for mental anomalism, i.e., the view that psychological explanation inextricably involves the representation of an agent's *reasons* for her actions, which in turn makes concepts involved in psychological explanation irreducibly normative: 'intentional action is action that can be explained in terms of beliefs and desires whose propositional contents rationalize the action' (Davidson 1982: 97). This is the source of Apperly's (2011) view that full-blown mindreading is *normative*. The normativist construal of mindreading reflects the assumption that mindreading is not only necessary but also sufficient for representing an agent's reasons. While full-blown mindreading is clearly meta-representational, one may resist the idea that it is also intrinsically normative.

For the purpose of predicting an agent's instrumental action (e.g., retrieving her toy), it may be necessary and sufficient to represent the contents of her desire and epistemic state, including the content of her false belief if she is mistaken. For example, in order to predict a mistaken agent's instrumental action, it is sufficient to know where she placed her toy before someone else moved it elsewhere in the agent's absence. While it is necessary to represent the content of her false belief if she is mistaken, it may not be necessary to assess it *as false*. However, for the purpose of explaining, justifying, appraising, or criticizing an agent's instrumental action, it is not only necessary to be able to assess a mistaken agent's belief as false, but also to represent her reasons. An agent's *objective* reason for looking for her toy at a location comprises her desire for the toy and the fact about the toy's location. If an agent holds a false belief about the location of her toy, then she will also have a *subjective* reason not to look for her toy at its actual location, but at some other empty location. If so, then she will fail to find her toy. So her subjective reason will fail to match her *objective* reason, which is to look for her toy at its actual location. Clearly, one could not represent the difference between an agent's mistaken subjective reason and her objective reason, let alone either explain the failure of her action or try to convince her that she should revise the content of her belief, unless one had the capacity to assess her false belief as false. The capacity to meta-represent the contents of an agent's beliefs and desires is a necessary

component of the capacity to represent and evaluate the agent's reasons, but it is far from clear that it is also a *sufficient* condition.

Nevertheless, many psychologists are inclined to think that full-blown mindreading is both necessary and sufficient for representing an agent's reasons, including the distinction between her objective and her subjective reasons, if they diverge. In particular, Perner and Roessler (2010, 2012) and Roessler and Perner (2013) have recently argued that preschoolers fail explicit false-belief tasks about an object's location because they fail to understand the divergence between the mistaken agent's objective reason for looking for her toy at its actual location and her subjective reason for looking for it at the empty location. They further assume that this failure demonstrates the fact that preschoolers lack full-blown (i.e. explicit) mindreading capacities. In Perner and Roessler's terminology, preschoolers have an implicit theory-of-mind, but they lack an explicit theory-of-mind.[6]

I now turn to confirmation holism as a way of explaining why the representation of the contents of others' beliefs may be considered effortful. Advocates of the two-systems model of mindreading accept Fodor's (1983) distinction between modular processes, which are informationally encapsulated, and the central processes underlying belief fixation, which Fodor takes to be isotropic and subject to confirmation holism. While the processes underlying minimal mindreading are taken to be informationally encapsulated, full-blown mindreading is taken to be subject to confirmation holism: if belief fixation is subject to confirmation holism, then a fortiori so is the fixation of beliefs about the contents of others' beliefs (cf. Apperly 2011).[7]

Now Fodor's own approach to the fixation of beliefs rests on his assumption that the confirmation of scientific hypotheses is our best model for the process of belief fixation and also on his joint acceptance of Quine's (1953, 1960) holistic view of scientific confirmation. Quine's own agenda was to use confirmation holism as a step in his argument for the revisability of logical laws and against the analytic–synthetic distinction. Arguably, confirmation holism makes the process of belief fixation—of either scientific or non-scientific beliefs—puzzling. But on the one hand, the fixation of beliefs about others' beliefs should be taken to be subject to confirmation holism exactly to the same extent that the fixation of beliefs about any other topic is. On the other hand, confirmation holism has never prevented either scientists or non-scientists to fix their beliefs. Where

[6] For another illustration of the widespread assumption that full-blown mindreading encompasses the capacity to attend to reasons, see the reaction by Low et al. (2016) to findings reported by Scott et al. (2015). The findings by Scott et al. (2015) suggest that seventeen-month-olds understand a thief's intention to covertly cause the owner of rattling toys to have false beliefs about her toys. Low et al. (2016) object to Scott and colleagues' mentalistic interpretation of their findings on the grounds that it would commit the infants to tolerating a high level of irrationality on the part of the thief.

[7] Thus, some philosophers (e.g., Zawidzki 2013) have further argued that acceptance of confirmation holism makes mindreading 'computationally intractable'.

one philosopher sees the opportunity for *modus ponens*, another may see the opportunity for *modus tollens*: neither Fodor's assumption that scientific confirmation is our best model for the fixation of beliefs in general nor Quine's holistic view of scientific confirmation is immune to doubt.

Finally, Frege (1892) famously appealed to the aspectuality of beliefs to resolve one version of his identity puzzle: how could the truth of (1) and (2) fail to entail the truth of (3)?

(1) Marta believes that the evening star is shining.
(2) The evening star = the morning star.
(3) Marta believes that the morning star is shining.

(3) is the output of the replacement of 'the evening star' by 'the morning star' in (1), licensed by the truth of the identity claim (2). If sentence (1) was an extensional context, then the truth of (1) and (2) would entail the truth of (3). On a *de dicto* reading of (3), the truth of (3) does not follow from the truth of (1) and (2).[8] But on a *de re* reading of (3), the truth of (1) and (2) does entail the truth of (3). The fact that on a *de dicto* reading of (3), the truth of (1) and (2) does not entail the truth of (3) is evidence that sentence (1) is *intensional*, not extensional. The intensionality of belief report (1) reflects the aspectuality of Marta's belief that the evening star is shining. It is evidence that the particular *way* the content of Marta's belief is being characterized, namely as the belief that the evening star (not the morning star) is shining, matters to the truth of the belief ascription. In other words, Marta can be a rational person and take two different attitudes with respect to the propositions expressed respectively by 'the evening star is shining' and 'the morning star is shining': she may hold the first true and the second false, because the propositions are different. This is Frege's solution to one of the versions of his puzzle about identity.

The aspectuality of beliefs (and other propositional attitudes) is widely regarded as a reliable sign of the propositional character of their contents. The aspectuality of thoughts and other propositional attitudes is one of the premises used by Davidson (1982) to argue for the thesis that non-human animals cannot think.[9] Understanding the aspectuality of others' beliefs is correspondingly widely

[8] On its *de dicto* reading, a belief ascription aims at capturing the way the believer would express the content of her belief. On its *de re* reading, a belief ascription relies on what is common ground between a speaker and his audience without taking into account the believer's own perspective.

[9] Davidson's thesis seems to lie in the background of Hutto's (2008) social enactivist claim that infants and non-human animals, who cannot entertain genuine propositional attitudes (because they do not speak a natural language), may nonetheless entertain what he calls 'intentional attitudes', which he takes to 'involve a kind of intentional directedness which is not semantically contentful'. Hutto further claims that correct descriptions (or attributions) of relevant instances of intentional directedness, which lack genuine semantic content, are extensional, not intensional. Zawidzki (2013) argues that the holism of belief confirmation that is taken to generate the intractability of mindreading reflects the aspectuality of propositional attitudes.

regarded as a demanding psychological task. The aspectuality of beliefs is mirrored by the intensionality (or referential opacity) of linguistic belief reports. The intensionality of the *de dicto* reading of a belief report stands in sharp contrast with the extensionality of the representation of an individual's behaviour (behaviour-reading), which in turn reflects the relational (non-aspectual) character of the individual's behaviour (e.g. kicking or pushing).[10]

Advocates of the two-systems model propose that while representing the aspectuality of beliefs requires the flexibility of full-blown mindreading, representing others' *registrations* can be achieved by minimal mindreading.[11] They take registration to be a non-aspectual epistemic relation between an agent, an object, and a location. On this account, representations of the registration relation are *extensional* (not intensional), as illustrated by the following pattern of inference:

(4) Marta registers <evening star, sky>
(5) The evening star = the morning star
(6) Marta registers <morning star, sky>

Butterfill and Apperly stipulate that the truth or correctness of (4) and (5) entails the truth or correctness of (6) (2014: 622). Thus, the way Marta registers the presence of the evening star in the sky (Marta registers <evening star, sky>) does *not* matter to the correctness or truth of the representation of the registration relation.

In a nutshell, the basic claim made by advocates of the two-systems model is that the representation of another's registration (which they construe as the representation of a genuine *non-propositional* epistemic state) constitutes the middle ground between the representation of another's belief *as such* and behaviour-reading. Only the full-blown (i.e. flexible, less efficient, later developing) mindreading system can represent the aspectual contents of others' beliefs and can thereby represent the contents of others' beliefs *as such*. The minimal (i.e. efficient, inflexible, earlier developing) mindreading system can track the contents of others' beliefs, *not* by representing them *as such*, but by representing the contents of others' registrations.

4. The Developmental Puzzle

Recent experimental developmental investigations of false-belief understanding in human children fall into six broad categories of false-belief tasks, according to

[10] If Marta pushes Bill and if Bill is Bob's father, then Marta pushes Bob's father.

[11] To the extent that success in so-called Level-2 visual perspective-taking tasks requires understanding the aspectuality of others' visual epistemic states, advocates of the two-systems model are also committed to the claim that Level-2 visual perspective-taking tasks can only be achieved by the full-blown mindreading system, not by minimal mindreading.

whether false-belief understanding is measured by means of an *explicit* (verbal) or an *implicit* (non-verbal) test. In most fully explicit tests, participants are directly asked a question by the experimenter. Most implicit or so-called 'spontaneous-response' tests use participants' looking behaviour in either the violation-of-expectation or the anticipatory gaze methodology. These experiments involve so-called familiarization (or habituation) trials whose purpose is to generate expectations in participants. Experiments based on the violation-of-expectation further involve test trials that may either be congruent or incongruent with the participants' expectations and the experimenters measure participants' looking *time* in response to respectively congruent and incongruent test trials. In experiments based on anticipatory gaze, the experimenters code the *location* of participants' first saccade in anticipation of the agent's action. Some experiments stand somewhere in between fully explicit and fully implicit measures, as when participants are encouraged to help a mistaken agent achieve the goal of her instrumental action.

Most recent experimental developmental investigations of false-belief understanding have been devoted to change-of-location false-belief tasks. In such tasks, participants see an agent place some toy in one of two opaque containers. In her absence, the toy's location is switched. The question is whether participants, who know the toy's actual location, can represent the content of the agent's false belief about it. (In so-called low-inhibition tasks, the toy simply disappears so that participants don't know its location. In such cases the participants' own knowledge of the toy's location cannot interfere with their representation of the content of the mistaken agent's false belief).[12]

The recent developmental psychological investigation of mindreading has given rise to discrepant findings, thereby generating a significant empirical puzzle. On the one hand, solid evidence shows that not until they are at least four-and-a-half years old can the majority of human children pass explicit or elicited-response false-belief tasks of various sorts, in which they are directly asked a question. For example, in the famous explicit Sally-Anne task, participants who know the toy's actual location are asked to predict where Sally (the mistaken agent) is likely to look for her toy or where she thinks her toy is. Most three-year-olds fail the task and point to the toy's actual location (Wimmer and Perner 1983; Baron Cohen et al. 1985; Wellman et al. 2001).

However, consider Onishi and Baillargeon's (2005) deservedly famous study. In their familiarization trials, fifteen-month-olds saw an agent that provided behavioural evidence that she was motivated to play with a toy (a watermelon), which

[12] Other false-belief tasks include unexpected content tasks, in which participants are shown an opaque box whose content is unexpected (e.g., a smarties box that contains crayons). The question is whether young participants are able to represent the content of their own previous false belief, or the potential content of another agent's false belief, about what is inside the box. Still other tasks probe participants' ability to represent the content of another's false belief either about the identity of a single object with two aspects or about the identity of two indistinguishable objects.

she placed into a green opaque box located next to a yellow opaque box. In four different belief-induction trials, the infants saw the toy either move from the green to the yellow box or not, in either the presence or the absence of the agent. Then in the test trials, they saw the agent reach for either the green or the yellow box. Onishi and Baillargeon found that infants looked reliably longer when the agent reached for the empty location with a true rather than a false belief and also when she reached for the correct location with a false rather than true belief. In a study by Buttelmann et al. (2009), a first experimenter placed her toy in one of two opaque containers and then left. In her absence, a second experimenter moved the toy from the first to the other container. In the true-belief condition, everything was the same except that the first experimenter was present when the second experimenter moved the toy. Finally, the first experimenter returned and tried unsuccessfully to open the container in which she had initially placed her toy. In the false-belief condition (but not in the true-belief condition), when they were invited to help the mistaken agent, eighteen-month-olds opened the container that contained the first experimenter's toy. Furthermore, Setoh et al. (2016) have recently reported evidence that two-and-a-half-year-olds succeed on an explicit low-inhibition change-of-location false-belief task.

Studies devoted to false-belief understanding about unexpected contents also exhibit a dissociation between findings based on respectively explicit and implicit tasks. Most explicit studies have shown that when asked what they earlier thought or what another would think when first confronted with the appearance of a smarties box, which in fact contains crayons, most three-year-olds incorrectly answer that they thought, and that someone else would think, that it contains crayons (cf. Perner et al. 1987; Gopnik and Astington 1988). However, two studies have recently shown that younger children can represent the contents of others' false-beliefs about unexpected contents. For example, He et al. (2011) have reported evidence based on the violation-of-expectation paradigm that two-and-a-half-year-olds look reliably longer when an agent reaches either for crayons in a Cheerios box or for Cheerios in a crayon box, after the contents of the boxes have been switched in the agent's absence, but not if the agent was present. In a study based on the helping paradigm, Buttelmann et al. (2014) familiarized eighteen-month-olds with boxes for blocks that contained blocks. When they subsequently saw an experimenter unsuccessfully reach for a box for blocks which they knew to contain spoons, infants based their choice of whether to helpfully give a spoon or a block to the experimenter on whether she had a true or a false belief about what was inside the block box.

In short, recent research yields discrepant developmental findings. The basic developmental puzzle is: why do three-year-olds fail explicit change-of-location or unexpected-contents false-belief tasks if toddlers or even preverbal infants can represent the contents of others' false beliefs about either an object's location

or unexpected contents? Until recently, there were two main responses to this developmental puzzle. Advocates of cultural constructivist approaches to mindreading, who assume that only success on explicit false-belief tasks could be evidence of false-belief understanding, argue that preverbal infants cannot understand the contents of others' false beliefs. Their task is therefore to offer low-level deflationary, entirely non-mentalistic accounts of the data consistent with their assumption that infants are unable to represent the contents of others' false beliefs. Some have argued for low-level associationist accounts and also for behaviour-reading heuristics (Perner and Ruffman 2005), others for sub-mentalizing processes involving perceptual novelty and retroactive interference (Heyes 2014).[13]

Critics of cultural constructivism subscribe to a nativist view of mindreading. Their burden is to explain why explicit change-of-location false-belief tasks are so challenging for most human children until they are four-and-a-half years old. Advocates of the processing-load account (Baillargeon et al. 2010) have argued that success in explicit tasks rests on three separable processes: (i) the representation of the content of the agent's false belief; (ii) a response-selection process whereby participants understand the relevance of the agent's false belief to the question asked; and (iii) a response-inhibition process whereby participants must inhibit any prepotent tendency to answer the test question based on their own knowledge of the toy's location (Baillargeon et al. 2010; Carruthers 2013). They argue that until they are four-and-a-half-years old, most children are overwhelmed by the demands of these three processes. In particular, they are taken to lack the executive resources required for achieving (iii) the response-inhibition process.

More recently, some critics of the cultural constructivist approach have also argued for a pragmatic approach to the developmental puzzle, according to which young children fail explicit false-belief tasks, not because they cannot represent the content of the agent's false belief, but because the question asked by the experimenter is pragmatically misleading (cf. Helming et al. 2014; 2016; Westra 2017a; Westra and Carruthers 2017). What lies at the root of the pragmatic approach is the fact that knowing where a mistaken agent placed her toy is sufficient either for predicting where she will look for it or for knowing where she thinks her toy is. However, as stressed by Helming et al. (2014, 2016), in explicit false-belief tasks, participants are confronted with two separate actions: the instrumental action performed by a mistaken agent and the communicative action performed by the experimenter. Helming and colleagues further argue that since success on explicit tasks requires participants to take a third-person perspective on the mistaken agent's instrumental action and a second-person perspective on the experimenter's communicative action, young children may be overwhelmed by this perspectival conflict.

[13] Cf. Carruthers (2013), Helming et al. (2014) and Jacob (forthcoming, 2018) for detailed criticisms.

Moreover, in classical scenarios of explicit change-of-location false-belief tasks, participants are further provided by the experimenter with much irrelevant information about the location of the object sought by the mistaken agent. While this information is strictly speaking irrelevant to *predicting* where the agent will look for her toy, it is nonetheless relevant to the *normative* evaluation of the agent's failure to achieve the goal of her instrumental action, which is to satisfy her desire to find her toy. If this is correct, then as suggested in recent papers by Perner and Roessler (Perner and Roessler 2010, 2012; Roessler and Perner 2013), the proper normative evaluation of the agent's failure to achieve her goal may require the normative distinction between an agent's *objective* reason and her *subjective* reason for her action. The mistaken agent has an *objective* reason to look for her toy at its *actual* location. But given her false belief about the toy's location, she has a *subjective* reason to look for it at the *empty* location. Not until they are comfortable with this normative distinction are young children likely to succeed on explicit false-belief tasks about the object's location. One interesting point of contention is whether, as Perner and Roessler have argued, it is the job of the mindreading capacity proper, not merely to accurately represent the contents and attitudes of others' mental states, but also to represent the normative distinction between an agent's objective reason and her subjective reason.

5. Can the Two-Systems Model Resolve the Developmental Puzzle?

It is one of the fundamental motivations of the two-systems model to offer a novel *middle ground* solution to the developmental puzzle presented in section three, which is different from both the cultural constructivist and the nativist approaches (cf. Apperly 2011; Apperly and Butterfill 2009; Butterfill and Apperly 2014; Low et al. 2016). According to advocates of cultural constructivism, only explicit change-of-location false-belief tasks can genuinely probe false-belief understanding. Nativists argue that the data based on implicit false-belief tasks show false-belief understanding in preverbal infants. The two-systems' middle ground approach to the developmental puzzle rests on the basic assumption that while the early-developing efficient and inflexible system of mindreading enables preverbal infants to represent others' true and false *registrations*, only the more flexible later-developing system enables human adults and older children to represent the aspectuality of beliefs *as such*.

According to the two-systems model, the early-developing efficient minimal mindreading system enables infants to represent the contents of others' false registrations about objects' locations, which explains the infants' data, based on implicit tasks. But only when the later-developing more flexible full-blown mindreading system is in place can most four-and-a-half-year-olds pass explicit

change-of-location false-belief tasks, which explains why most three-year-olds fail explicit change-of-location false-belief tasks.

To the extent that registration is construed as a ternary relation between an agent, an object, and the object's location, the content of an agent's registration seems to be the relational (non-propositional) counterpart to the propositional content of an agent's belief about an object's location. If so, then the ability to represent the relational content of an agent's registration seems to be the *minimal* counterpart to the ability to represent the propositional content of an agent's belief about an object's location.

In a nutshell, a minimal mindreader can represent the contents of others' registrations, not the contents of others' genuine beliefs (even about an object's location). Thus, minimal mindreading purports to stand as a tentative middle ground between the nativist and the cultural constructivist approaches to the ontogenetic development of mindreading capacities in humans. Whether minimal mindreading does indeed constitute a stable middle ground position is an open and delicate question. Unlike advocates of the nativist approach, the two-systems model denies that the capacity to represent the contents of others' beliefs as such is present in human infancy (and could thereby be innate). On the two-systems model, only the capacity to represent the contents of others' registrations is present in human infancy (and could thereby be innate). On this model, children presumably bootstrap their way to full-blown mindreading on the basis of minimal mindreading and language acquisition. However, it is a delicate issue whether and to what extent minimal mindreading is a genuine alternative to such versions of cultural constructivist approaches to infant mindreading as behaviour reading, associationism, and sub-mentalizing. Arguably, an agent's registration of an object at a location should be confused neither with the agent's behaviour strictly speaking nor with perceptible features of the agent's non-mental environment, in accordance with the perceptual novelty approach recommended by advocates of the sub-mentalizing approach. However, to the extent that an agent's registration of an object at a location is construed as an *extensional* non-mentalistic relation, it suspiciously looks like a ternary association between an agent, an object, and a location.

The two-systems model rests on the split between epistemic states that have and those that do not have minimal counterparts: beliefs about an object's location do, but beliefs about object-identity do not, have minimal counterparts. Minimal mindreaders can track the contents of others' true and false epistemic states about *objects' locations without* representing them *as such*; but they can't track the contents of others' true and false epistemic states about *object-identity without* representing them *as such*.[14] Can minimal mindreaders also track the contents

[14] On the two-systems model, minimal mindreaders can represent the contents of others' registrations, but not the contents of others' beliefs as such. It so happens that the contents of others' registrations overlap with the contents of others' beliefs about an object's location as such. Minimal mindreaders lack

of others' true and false epistemic states about *unexpected contents without* representing them *as such*? It is unclear how the two-systems model should answer this question.

I now want to argue that the two-systems model faces three basic challenges. First, the notion of registration fails to meet a condition of adequacy that is built into the two-systems model. Secondly, I want to suggest that the two-systems model does not really resolve the developmental puzzle. Finally, I want to argue that registration might not be sufficient for handling the basic findings about infants based on implicit change-of-location false-belief tasks.

I turn to the first question first. According to Apperly and Butterfill's official definition, 'one stands in the registering relation to an object and location if one encountered it at that location and if one has not since encountered it somewhere else' (2009: 962). So the notion of *registration* is defined in terms of the notion of *encountering*. An agent is further said to stand in the encountering relation to an object if the object stood in the agent's *field* at a given instant and was not visually occluded from the agent's line of sight (or otherwise blocked from the agent's sensory processing) (see Butterfill and Apperly 2014: 614). They construe encountering and registration as 'non-representational proxies for perception and belief' (2014: 624). As they strongly emphasize, encountering is a *non-aspectual* relation between an agent, an object, and a location. Given the definition of registration in terms of encountering, registration is expected to inherit its non-aspectuality (or relational character) from the non-aspectuality of the encountering relation. Representing an agent's registration of an object at a location should be extensional to the same extent that representing the agent's encounter with that object is extensional.

Clearly, the condition of adequacy that is built into the two-systems model, and that the notion of registration ought to satisfy, is that representing the contents of others' registrations (which is achieved by the early-developing efficient system) should be cognitively easier and less demanding than representing the contents of others' beliefs (which can only be achieved by the later-developing more flexible system). The very fact that the contents of others' registrations are non-aspectual and that the representations of others' registrations are purely extensional is further evidence that representing the contents of others' registrations fits this condition.

But consider what the official definition implies: an agent could not stand in the registration relation to an object and a location at time t unless the agent stood in the encountering relation to that object at some earlier time t−1 and she did not encounter the same object at *any* other location in the interval between t−1 and t. This entails in turn that one could only represent an agent's registration

therefore any means of tracking the contents of others' beliefs about other topics than an object's location. In particular, they are unable to track the contents of others' false beliefs about object-identity.

of an object at a location at t if (i) one could represent the agent's encountering the object at the same location at t−1 and (ii) one could further represent the fact that the agent failed to encounter the same object at *any* other location in the interval between t−1 and t. Condition (ii) is important because it specifies the extent to which representing a registration goes beyond representing an encounter. But if so, then only if one could represent the *negation* of the encountering relation and also *universally quantify* over places could one represent another's registration of an object at a location. So the question arises: to what extent does registration satisfy the condition of adequacy according to which representing others' registrations should be significantly less demanding than representing others' beliefs?

Several critics (including two referees for this paper) have suggested two lines of defence on behalf of the two-systems model, the first of which is that it is one thing to assume that an agent stands in the registration relation to an object at a location if she stood in the encountering relation with this object at the same location at an earlier time and did not encounter the object at a different location after this time. It is quite another thing to further claim, as I do, that one could not represent the agent's registration relation with an object and a location unless one could represent both the fact that the agent earlier stood in the encountering relation with the same object at an earlier time at the same location and the fact that the agent did not encounter this object elsewhere at a later time. For example, these critics point out that from the fact that content externalists claim that an individual could not think about water unless she stood in the causal relation to water, it does not follow that content externalists are thereby committed to the view that an individual could not think about water unless she could also represent the causal relation between water and herself. Similarly, many epistemologists assume that the mere lack of relevant alternatives to the truth of a proposition *p* is a sufficient condition for an agent to know proposition *p*. It is not further necessary that the agent also be able to represent the lack of relevant alternatives in order to know that *p*. I do agree that it is not necessary for an agent *to stand* in the registration relation to an object and a location at a time that she knows (or represents) both the fact that she earlier stood in the encountering relation to the same object and location and the fact that she did not encounter it elsewhere at some later time. But I do maintain that given the two-systems model's definition of registration, a minimal mindreader could not *represent* the fact that an agent stands in the registration relation to an object and a location unless she could also *represent* the fact that she earlier encountered the same object at the same location and failed to encounter it elsewhere since then.

A second line of defence suggested by a referee for this paper is that if the ability to represent the contents of others' registrations does require, as I argue, the ability to use negation (of the encountering relation) and universal quantification (over places), then this may well be consistent with the purported informational encapsulation of the early-developing system. Perhaps this is correct. If so, then

advocates of the two-systems model can happily assume that the early-developing system is informationally encapsulated and also requires the ability to use negation and universal quantification. But it is worth pointing out in this context that in her impressive study on concepts, Carey (2009) argues that the ability to use negation and universal quantification rests on language acquisition.

I now want to cast doubt on the claim that the two-systems model can resolve the puzzle of the discrepant developmental findings: why do three-year-olds fail explicit change-of-location false-belief tasks if findings based on implicit tasks show that preverbal infants expect agents to act in accordance with the contents of their true and false beliefs? Now the typical structure of the mistaken agent's instrumental action in implicit change-of-location false-belief tasks is exactly the same as the structure of the mistaken agent's instrumental action in explicit change-of-location false-belief tasks (e.g., the Sally-Anne task).[15] In either implicit or explicit tasks, a mistaken agent places her toy in one of a pair of opaque containers and in her absence the toy's location is switched. But, as advocates of the two-systems model have claimed, the ability to represent true and false registrations is sufficient to account for such findings as Onishi and Baillargeon's, which 'could be explained on the hypothesis that [infants] are tracking registration as a cause of action' (Butterfill and Apperly, 2014: 620).[16]

The fact that the structure of a mistaken agent's instrumental action is the same in both implicit and explicit change-of-location false-belief tasks is fundamental for addressing the question whether the two-systems model can resolve the developmental puzzle. If representing the content of another's registration is sufficient to account for infants' responses in implicit change-of-location false-belief tasks and if the false-belief scenario is the same whether the task is explicit or implicit, then it should also be sufficient for securing participants' understanding of the content of the mistaken agent's false belief in explicit change-of-location false-belief tasks. Presumably, the only difference between an implicit and an explicit change-of-location false-belief task is that in the latter only, not in the former, participants are also directly asked an explicit question. If so, then presumably the advocate of the two-systems model should argue that success on explicit, but not on implicit, change-of-location false-belief tasks further requires the later-developing flexible system necessary for reading the experimenter's mind and recognizing her communicative intention. But if so, then advocates of the two-systems model seem committed to the following dilemma, neither horn of which should be very attractive to them. One option is that the early-developing and the later-developing

[15] Cf. Wimmer and Perner (1983); Baron-Cohen et al. (1985); Wellman et al. (2001).

[16] 'Registration also can be understood as determining which location an individual will direct their actions to when attempting to act on that object. This more sophisticated understanding (which requires the notion of an unsuccessful action) enables one to predict actions on the basis of incorrect registrations and so approximate belief reasoning to such a great extent as to pass some false-belief tasks (e.g., Onishi & Baillargeon 2005)' (Apperly and Butterfill 2009: 962–3).

systems of mindreading cooperate and speak to each other in order to explain success on explicit change-of-location false-belief tasks: the early-developing system is sufficient for tracking the content of the mistaken agent's false belief and the later-developing system is required for attributing a communicative intention to the experimenter. But this seems to contradict the fundamental assumptions made by advocates of the two-systems model that the two systems are separate, do not speak to each other, and that the early-developing system persists through adulthood alongside the later-developing system. The other option is that while the early-developing system is sufficient to resolve implicit change-of-location false-belief tasks, the later-developing system alone is involved in resolving explicit change-of-location false-belief tasks. In which case, the early-developing system must be inhibited either by the later-developing system itself or by some higher-level executive system. This option is also problematic for advocates of the two-systems model because the early-developing system is taken to be automatic and it is far from obvious how an automatic system could be inhibited, if at all.

So far, I have assumed, along with advocates of the two-systems model, that representing the contents of others' registrations is indeed sufficient for explaining the infants' data based on implicit change-of-location false-belief tasks. Now, I want to cast doubt on this assumption. The problem is that registration is officially defined as a ternary *unstructured* relation between an agent, a toy and a location: (R<agent, toy, location>). On this official unstructured relational construal, an agent can register the presence of a toy *at* a location. Now consider the experimental design of Onishi and Baillargeon's (2005) test trials: each of the pair of visible green and yellow opaque boxes is *at* a location. However, what matters to the agent is the *invisible* toy, which happens to be *within* (or *inside*) one of the pair of boxes (e.g., the green box), not the pair of coloured visible boxes, each of which is at a location. Unless the advocates of the two-systems model were to endorse one of the non-mentalistic accounts of such implicit change-of-location false-belief tasks along the lines of either Perner and Ruffman's (2005) associationist proposal or Heyes's (2014) sub-mentalizing proposal, advocates of the two-systems model face the following dilemma: either registration is an *unstructured* ternary relation or it is not. If it is, then representing the content of the agent's registration is unlikely to be sufficient to enable infants to represent the content of the agent's true or false epistemic state about the location of her toy in the test trials of Onishi and Baillargeon's study. In the belief-induction trials, the agent last saw her toy being placed into one of the pair of boxes. However, in the test trials, the toy is invisible; each of the pair of boxes is at a location and one of the pair of boxes contains the invisible toy. So the invisible toy is inside or within one of the two boxes. In order to make sense of the agent's action of reaching into one of the boxes in the test trials, the infants must represent the fact that the toy is inside one of the pair of boxes. If registration is unstructured, then infants may represent the agent's registration of the box that contains the toy as being at its location, but not

the toy as being inside the box (which is at its location). If registration is not unstructured, then infants may represent the agent's registration of the toy as being inside the box which is at its location. But then it is likely that the content of the agent's registration is going to suspiciously look propositional, e.g. R<agent, within<toy, green box>>, in which relations are nested within one another. If so, then the gap between the contents of others' registrations and others' beliefs about an object's location becomes evanescent.

6. Is Aspectuality a Signature Limit of the Early-developing System?

One of the most interesting empirical claims made by advocates of the two-systems model is that representing the aspectuality of genuine beliefs (as displayed by the intensionality of belief reports on a *de dicto* reading) is beyond the limits of the early-developing efficient system. It is only within the capacities of the later-developing flexible system: representing the aspectuality of genuine beliefs should be a 'signature limit' of the early-developing system that could be displayed by adults as well as preverbal infants. However, this claim turns out to be disconfirmed by empirical evidence.

As I noted earlier, advocates of the two-systems model draw a sharp dichotomy between the cognitive challenges raised by change-of-location false-belief tasks and false-belief tasks about object-identity. However, as I also noted, there seems to be a continuum from change-of-location to object-identity tasks, ranging over unexpected-contents tasks. Registration is supposed to enable minimal mindreaders to track the contents of others' false beliefs about an object's location in implicit tasks. Full-blown mindreading is required for representing the contents of others' false beliefs about object-identity as such. However, the evidence shows that the developmental puzzle also arises for research based on object-identity false-belief tasks, which are widely taken to probe understanding of the aspectuality of beliefs.

Much evidence shows that *explicit* object-identity false-belief tasks are more challenging for young children than *explicit* change-of-location false-belief tasks. For example, in studies by Apperly and Robinson (1998; 2003), children between four and six years of age, who succeed on explicit change-of-location false-belief tasks, were shown two objects, one with a single aspect, the other one with two aspects: for example, one was an eraser and the other was an eraser that was also a die. The children were then introduced to an agent-puppet who only knew of the object with a dual nature that it was a die. The puppet was present when the eraser with a single aspect was placed into one opaque container and the object with a dual nature was placed into the other opaque container. When children were asked to predict in which of the two containers the puppet was likely to look for an eraser, they were at chance and selected at random between the locations of the two objects (see Rakoczy et al. 2015 and Perner et al. 2015 for discussion).

However, in a recent study, Rakoczy et al. (2015) suitably modified the above task by introducing a single object with a dual nature, which was both a die and an eraser. In the false-belief condition, only the children, not the puppet were informed of its dual nature; the puppet knew of the object only as an eraser. The object was placed into one of two opaque containers under its eraser aspect in the presence of both the children and the puppet. The children were reminded, in the puppet's absence, of the dual nature of the eraser. Finally, in the presence of both the children and the puppet, the object was moved under its die aspect from one container to the other. The children, who were the same age as in Apperly and Robinson's studies, were asked where the puppet would look for the eraser. Most children correctly pointed to the first container. In the true-belief condition, in which the puppet was aware of the dual nature of eraser-die, most children correctly pointed to the second container in response to the same question.

Furthermore, one additional study by Buttelmann et al. (2015) provides evidence that preverbal infants can represent the content of an agent's false belief about a single object with two distinct aspects. In this study, eighteen-month-olds were made aware that each of a set of target toys had a deceptive aspect: for example, a sponge that looked like a rock. For each of the set of target toys, the infants were provided with pairs of test objects, each member of which resembled either aspect of the target toys. Infants saw an agent who either knew about the two aspects of the target toys or did not, and who wanted to reach one of them but failed to grasp it. Infants were instructed to help the agent achieve her goal by giving her, not the very target object that the agent was unsuccessfully trying to grasp, but instead one of the pair of duplicate objects that were available to them. The infants reliably gave the agent the duplicate object that resembled the target under its aspect known to the agent, only in the false-belief condition (when the agent was not aware of the target's two aspects), not in the true-belief condition (when the agent was aware of the target's two aspects). This study strongly suggests that eighteen-month-olds are able to ascribe to someone else the false belief that there are two distinct objects when they know that there is a single object with two distinct aspects.[17]

In contradistinction to the findings by Buttelmann et al. (2015), Rakoczy (2017) reports the findings of a study investigating toddlers' understanding of aspectuality on the basis of the helping paradigm first used by Buttelmann et al. (2009) in the context of change-of-location false-belief tasks (discussed in section three). In this study, a toy with two aspects was placed in one of two boxes under one of its two aspects (aspect A) in the presence of both the agent and two-year-olds. But only the two-year-olds, not the agent, were aware of the toy's other aspect (aspect B). The toy was subsequently moved from the first to the second box under its B aspect

[17] Scott and Baillargeon (2009) report findings showing that eighteen-month-olds can also represent the content of an agent's false belief that two indistinguishable objects (e.g., a two-piece penguin and a one-piece penguin) are one and the same when the infants know that when the two-piece penguin is assembled it is perceptually indistinguishable from the one-piece penguin.

in the presence of both the agent and the infants. Since the agent was unaware of the toy's B aspect, she must have falsely believed that there were two objects, one in each box. Finally the agent unsuccessfully tried to open the first box, seemingly trying to retrieve the toy under its A aspect. The infants were invited to help the mistaken agent. Contrary to the findings reported by Buttelmann et al. (2009), Rakoczy and colleagues found that two-year-olds did *not* reliably help the mistaken agent by opening the second box that contained the single object with two aspects (no more so than in the true-belief condition). Rakoczy (2017) concludes that this finding vindicates the two-systems model's prediction that understanding the aspectuality of belief (or passing object-identity false-belief tasks) is beyond the limitations of the minimal mindreading system. This conclusion, however, is not inevitable. In order to efficiently help the mistaken agent find her toy, it is necessary that infants understand that the agent falsely believes that there are two toys (not one), one in the first box and another in the second box. But it is not sufficient. They could only efficiently help the mistaken agent find her toy by opening the second box if they further felt confident that the agent's *desire* to find her toy, which she only knows under its A aspect, would be fulfilled upon discovering the toy under its B aspect under which it was moved from the first to the second box. Lack of confidence about the fulfillment of the agent's desire might prevent two-year-olds from opening the second box.[18] If so, then it is not clear that this last finding offers support to the two-systems model.

In light of the study by Buttelmann et al. (2015) then, advocates of the two-systems model face the following dilemma, one horn of which is that the later-developing system (which is responsible for the representation of the contents of others' false beliefs about object-identity) is already present in eighteen-month-olds. The other horn of the dilemma is that the representation of the content of another's false belief about object-identity is not a signature limit of the early-developing system.[19]

[18] In the non-aspectual version of the helping study, if the agent is absent when the toy is moved from one location to the other, toddlers can confidently attribute a false belief to the agent. But this is not the case in the aspectual version of the helping study because the two aspects are enduring properties of the toy, which the agent might be aware of even if she is not present during the demonstration of the toy's two aspects. Consequently, an alternative interpretation of the fact that toddlers did not help the mistaken agent in the aspectual version of the helping study is that in the aspectual version they took the agent to know about the two aspects of the toy even when the agent was absent during the demonstration of the two aspects.

[19] In a pair of studies, Low and Watts (2013) and Low et al. (2014) present purported evidence for the claim that representing the content of another's belief about object-identity falls within the purview of full-blown mindreading, but beyond the resources of minimal mindreading. The authors probe the full-blown mindreading system by asking participants a direct question and the minimal mindreading system by coding participants' anticipatory gaze. They report that most adults reliably succeed on the explicit task and fail the implicit task. They argue that this dissociation is evidence for the two-systems model. However, there is an alternative explanation of the findings: success on one task requires belief-revision and success on the other task requires mental rotation (neither of which are specific to mindreading). While both belief-revision and mental rotation are time-demanding operations, there was no time limit in the explicit task; but participants' first saccade was coded only

7. The Cognitive Trade-Off Between Flexibility and Efficiency

As Apperly puts it, one of the central claims on behalf of the separation between two systems for mindreading is that:

> There is a tension between the requirement that mindreading be extremely flexible on the one hand, and fast and highly efficient on the other. Such characteristics tend not to co-occur in cognitive systems, because the very characteristics that make a cognitive process flexible—such as unrestricted access to the knowledge of the system—are the same characteristics that make cognitive processes slow and effortful. Instead, flexibility and efficiency tend to be traded against one another. (2013: 73–4)

In short, the bifurcation between minimal and full-blown mindreading systems rests to a large extent on the fundamental assumption that minimal mindreading is *efficient* because it is *automatic* (i.e., informationally encapsulated in Fodor's (1983) sense, cf. Figure 5.1). By contrast, full-blown mindreading is taken to be effortful, inefficient, flexible, and non-encapsulated (in Fodor's sense). As Butterfill and Apperly have put it, a process is *automatic* if it occurs *whether or not* it is relevant to participants' motives and goals (Butterfill and Apperly 2014: 608; cf. Carruthers 2015a; 2016).[20]

Putative evidence that human adults can achieve some mindreading tasks automatically in the relevant sense has been provided by several recent studies. For example, Kovacs et al. (2010) reported that adults whose psychophysical task was to press a button as fast as possible as soon as they detected a ball behind an occluder were faster when they expected the ball to be there rather than when neither they nor another agent (a blue smurf) expected it to be there. Kovacs and colleagues also found that participants were faster when they did not expect the ball to be there, but the blue smurf wrongly expected it to be there. Here it seems as if adults did compute the content of the blue smurf's false belief about the ball's location in spite of its irrelevance to their psychophysical task.[21] Van der Wel et al. (2014) further report that when participants reach toward a target object, the continuous trajectory of their reaching actions can also be modulated by the content of another's false belief.

1,750-ms after the extinction of the cue informing participants that the agent was about to act. Arguably, in the implicit task, participants did not have the time to perform either belief revision or mental rotation (cf. Carruthers 2015a; Jacob 2013, 2014).

[20] As two referees for this chapter note, it is an open question (which I leave entirely open) whether automaticity and speed always go together.

[21] In further work, Kovacs et al. (2014)Click here to enter text. have construed their earlier findings in terms of *spontaneous* rather than strictly automatic processes, where a spontaneous process is one not triggered by external instructions such as an experimenter's request.

By contrast, Apperly et al. (2006) and Back and Apperly (2010) report putative evidence that the representation of the contents of others' false beliefs is effortful and not automatic. In these studies, participants see a male agent hide a ball under one or another cup either in the presence or the absence of a female agent. Participants are explicitly instructed to track the location of the ball. Occasionally they are unexpectedly probed about their representation of the content of the female agent's true or false belief (about the ball's location). Apperly et al. (2006) and Back and Apperly (2010) found that participants' reports about the female agent's beliefs are slower than their reports about the ball's location. This temporal difference vanishes if they are explicitly instructed instead to keep track of the female agent's beliefs. This evidence is compatible with the non-automaticity of mindreading. However, Cohen and German (2009) modified the above experimental design by significantly shortening the temporal interval between the event whereby the female agent formed her (true or false) belief and the event whereby participants were probed for reporting their representation of the agent's belief. Under such a modification, the mindreading task is less taxing for participants' working memory. Cohen and German found that participants' responses about the agent's beliefs were just as fast as their responses about the location of the ball. Thus, whether the process whereby adults represent the contents of others' true and false beliefs is automatic or not is an open question.

In several further studies, Apperly and colleagues have systematically investigated some of the contrasts between Level-1 and Level-2 visual perspective-taking tasks (cf. Flavell et al. 1981). In Level-1 visual perspective-taking tasks, participants are requested to understand that two agents may not see the same things because some things which are visible to one may be occluded to the other. For example, in a 'dot-perspective' study of Level-1 visual perspective-taking by Samson et al. (2010), adults were asked how many dots they could see in a scene displaying a room with three walls and an avatar facing one of the walls. Samson and colleagues found the following *altercentric* effect: participants' answers were faster and more accurate when they saw the same number of dots as the avatar rather than when they did not. This suggests that participants automatically computed the avatar's Level-1 visual perspective despite the fact that it was not relevant to answering the question of how many dots they could see.[22]

In Level-2 visual perspective-taking tasks, participants are requested to understand that two agents may see one and the same thing differently or under distinct aspects. (To some extent, Level-2 visual perspective tasks probe participants' understanding of the aspectuality of others' epistemic visual perceptual states.)

[22] Santebastian et al. (2014) reproduce the altercentric effects of the 'dot-perspective' by replacing the avatar with an arrow. They argue that this is evidence that the dot-perspective task can be achieved by domain-general processes of sub-mentalizing. However, arrows may be interpreted by participants as symbols used by (and therefore as proxies for) agents with beliefs and desires for the purpose of providing information.

For example, in Surtees et al.'s (2012) 'numeral-perspective' study involving Level-2 visual perspective-taking, participants saw an avatar facing them, sitting at a table on top of which a numeral was displayed. On consistent trials, participants and the avatar could see the numeral displayed on the table in the same way (e.g., '8'). On inconsistent trials, participants and the avatar could not see the numeral displayed on the table in the same way (e.g., '6'). Surtees et al. (2012) did not find any *altercentric* effect: when asked what they could see, participants were not slower nor less accurate in the inconsistent than in the consistent trials. This suggests that participants did not automatically compute the avatar's Level-2 visual perspective.

Thus, there is some empirical support for the joint claims that the automatic execution of Level-1 visual perspective-taking tasks by human adults directly reflects the efficiency and the encapsulation of minimal mindreading, but the effortful execution of Level-2 visual perspective-taking tasks by human adults directly reflects the distinctive flexibility and non-encapsulation of full-blown mindreading. However, both claims have recently been subjected to interesting criticisms by advocates of the one-system approach to mindreading (in particular, Carruthers 2015a, 2016; and Westra 2017b).

I start with the claim that Level-2 visual perspective-taking tasks require effortful cognitive resources that are intrinsic to the mindreading capacities. First of all, as Carruthers (2015a, 2016) has argued, in order to judge whether the avatar sees a stimulus as '6' or '9' (in Surtees et al.'s Level-2 perspective-taking task), when participants themselves see it as '6' they need to take their own mental image of the stimulus and mentally rotate it until it matches the avatar's spatial position and orientation. This mental rotation requires executive working memory resources necessary to sustain and manipulate one's visual representation. But the process of mental rotation itself is part of mental visual imagery, not mindreading. If this is correct, then Level-2 visual perspective-taking tasks are effortful, not because mindreading is effortful, but because mental rotation is.

Secondly, as Carruthers (2015a; 2016) has further argued, in the 'numeral-perspective' study by Surtees et al. (2012), participants might lack sufficient *motivation* for representing a mere *avatar*'s visual perspective onto a numeral. If so, then they might not feel the urge to engage their working memory resources necessary for performing the task of mentally rotating their own visual image of the numeral in order that it matches the avatar's spatial orientation. This would explain the absence of altercentric effects (which are taken to be a signature of Level-1 visual perspective-taking by Apperly and colleagues).

To a large extent, Carruthers' motivational diagnosis is corroborated by a recent study by Elekes et al. (2016), who used a modified version of Surtees et al.'s (2012) numeral-perspective task. All participants perform a number verification task in which they check whether a numeral which they can see on a computer screen lying flat in front of them does or not represent the same number as a number word heard from an audio recording. Elekes and colleagues compared three

conditions: in the individual (or baseline) condition, participants perform the number verification task alone. In a pair of joint conditions, participants face a live partner while they perform the number verification task. In the joint perspective-dependent condition, they are informed that their partner is completing the very same task (and therefore shares their goal, even if they are not involved in performing a joint action). In the joint non-perspective-dependent condition, they are told that their partner is completing a different task, e.g., judging whether the colour of the numeral being presently displayed on their computer screen is the same as the colour of the numeral previously displayed on their computer screen. Elekes and colleagues report an altercentric effect (i.e., participants were slowed down) in the joint condition relative to the individual condition, but only if they knew or believed that their partner shared their own goal and their partner's response would diverge from their own on the basis of Level-2 visual perspective-taking (e.g. '6' or '9', not '0' or '8').[23] As Westra notes, these findings show that 'Level-2 perspective-taking can, at times, be fast and efficient, if subjects are provided with the right background knowledge and are sufficiently motivated. This contradicts the claim that Level-2 perspective-taking is a slow and effortful process' (Westra 2017b: 4572).

I now turn to the claims that Level-1 visual perspective-taking tasks can be automatically executed by human adults and that this automaticity directly reflects the efficiency and the encapsulation of minimal mindreading. In Level-1 visual perspective-taking tasks, participants must determine *what* another agent is seeing. Closely related tasks have been investigated within the so-called 'gaze-cueing' paradigm, which rests on participants' ability to determine *where* another is looking. In typical gaze-cueing studies, participants see a human face in the centre of a screen whose eyes are looking straight at them but who can shift her gaze sideways to her left or to her right. Their task is to detect a target object that can appear on either side of the face in front of them. In congruent trials, the target object appears on the side towards which the face has just shifted her gaze. In incongruent trials, the target object appears on the alternative side. The general finding is that participants are *cued* by the individual's gaze shift: they are slower to detect the target object in the incongruent than in the congruent trials.

As insightfully noticed by Westra (2017b), several experiments show that participants' responses in the gaze-cueing paradigm are modulated by their background knowledge. For example, in a study by Teufel et al. (2010), participants first experienced either opaque or transparent goggles that looked exactly the same from the outside. In the gaze-cueing task, they all watched an agent whose face was looking straight at them, but who was wearing goggles. Only participants who had experienced transparent (not opaque) goggles were cued by the agent's head

[23] Surtees et al. (2012) report a very similar finding based on a set-up in which participants take turns with a partner, instead of performing the same task simultaneously.

movements. In other words, if participants thought that the agent could not see, then his head movements failed to cue them. Other experiments by Ristic and Kingstone (2005) show that when participants are presented with an ambiguous stimulus that can be construed as either a pair of eyes or a pair of wheels, they are only cued by what they take to be eyes, not wheels. Other studies have also shown gaze-cueing to be modulated by participants' knowledge of whether the agent whose face they see is an in- or an out-group member, whether they know his or her age, race, social status, and how threatening they perceive him or her to be. For example, if an agent is a low-status member of participants' in-group, then his or her face is less likely to gaze-cue participants than if he or she is a high-status member of participants' in-group or a threatening member of participants' out-group (cf. Chen and Zhao 2015).

It is plausible that if representing *what* an agent sees (Level-1 visual perspective-taking) is automatic, then so is representing *where* an agent is looking (gaze-cueing). Conversely it is equally plausible that if representing where an agent is looking is *not* automatic, then *neither* is representing what an agent is seeing. There is evidence that representing where an agent is looking is neither automatic not encapsulated from participants' background knowledge. This evidence undermines the claim that representing what an agent sees (Level-1 visual perspective-taking) is always automatic and encapsulated from participants' background knowledge. In short, what makes Level-2 visual perspective-taking tasks challenging may reflect demands of mental rotation, not mindreading per se. There is also evidence that enhancing participants' motivation may improve their performance in Level-2 visual perspective-taking tasks. If the presence of altercentric effects is taken as evidence of automaticity, then the findings by Elekes et al. (2016) should be taken as evidence that Level 2 visual perspective-taking can be automatic. Conversely, these findings might also be taken to cast doubt on the claim that the presence of altercentric effects should be taken as evidence of automaticity.

Finally, as Carruthers (2015a) and Westra (2017b) have argued, while the processes underlying Level-1 visual perspective-taking tasks turn out not to be automatic in all cases, the processes underlying Level-2 visual perspective-taking tasks turn out to be less effortful and more spontaneous than predicted by the two-systems model.[24] If so, then the contrast between Level-1 and Level-2 visual perspective-taking tasks cannot be mapped onto the distinction between *automatic* and *effortful* mindreading processes. Alternatively, one may draw a distinction between the spontaneous and the reflective usage of a single mindreading capacity. For example, children and adults alike have been shown to spontaneously ascribe to puppets and even to geometrical stimuli psychological states ranging from emotions to false beliefs. Adults, if not young children, are further able to suspend

[24] Although Butterfill and Apperly (2014: 608) acknowledge the distinction between automatic and spontaneous processes, they seem to miss its significance for their proposal.

their ascription on the reflective grounds that puppets and geometrical stimuli cannot have psychological states. As Carruthers (2015a) argues, a process may be construed as spontaneous to the extent that it is triggered by participants' endogenous (conscious or unconscious) goals, not exogenously triggered by another's instructions. Furthermore, a process may be more or less spontaneous: people may be spontaneously more *motivated* to read *some* minds than others, just as they may have more *background knowledge* about *some* minds than about others. All of this is consistent with a one-system approach to mindreading.

8. Conclusions

The two-systems model is a significant attempt at offering a middle ground approach to mindreading that stands half way between a cultural constructivist approach to mindreading and its nativist alternative. It aims at both resolving the developmental puzzle and making sense of data that suggest that adults perform some mindreading tasks automatically. In this chapter, I have highlighted three main critical points: (1) the two-systems model fails to offer a satisfactory resolution of the developmental puzzle; (2) the current evidence so far fails to vindicate the claim that the aspectuality of beliefs is a signature limit of the minimal mindreading system; and (3) the contrast between Level-1 and Level-2 visual perspective-taking tasks does not support the sharp distinction between the automaticity of one minimal mindreading system and the flexibility of a distinct full-blown mindreading system. In short, this chapter supports the picture of a single mindreading system that can be used in ways that are more or less effortful, as a result of its interactions with other cognitive systems, such as working memory, executive control, and pragmatic competence.[25]

[25] I am grateful to Anita Avramides, Will McNeill, and Matthew Parrott for their challenging questions, comments, criticisms, and editorial suggestions on this chapter. I am also grateful to Renée Baillargeon, Cristina Becchio, Steve Butterfill, Gergo Csibra, Gyuri Gergely, Bart Geurts, Katharina Helming, Celia Heyes, Dora Kampis, Agi Kovacs, Jennifer Nagel, Hannes Rakoczy, Paula Rubio-Fernández, Vicky Southgate, Dan Sperber, and Brent Strickland for discussions of topics addressed in this chapter. Finally I gratefully acknowledge support from the European Research Council (ERC) under the European Union Seventh Programme (FP/2007-2013)/ERC Grant 609819.